

128.

21 世纪统计学系列教材

抽 样 技 术

金勇进 蒋 妍 李序颖 编著



A0966697

中国人民大学出版社

图书在版编目 (CIP) 数据

抽样技术/金勇进等编著.

北京:中国人民大学出版社,2002

21 世纪统计学系列教材

ISBN 7-300-04079-9/F·1254

I. 抽…

II. 金…

III. 抽样调查—高等学校—教材

IV. C811

中国版本图书馆 CIP 数据核字

21 世纪统计学系列教材

抽样技术

金勇进 蒋 妍 李序颖 编著

出版发行:中国人民大学出版社

(北京中关村大街 31 号 邮编 100080)

邮购部:62515351 门市部:62514148

总编室:62511242 出版部:62511239

E-mail:rendafx@public3.bta.net.cn

经 销:新华书店

印 刷:北京金特印刷厂

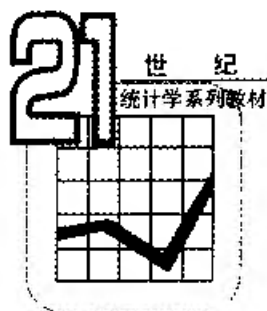
开本:787×965 毫米 1/16 印张:23

2002 年 6 月第 1 版 2002 年 6 月第 1 次印刷

字数:420 000

定价:28.00 元

(图书出现印装问题,本社负责调换)



总 序

中国人民大学应用统计科学研究中心 2000 年 6 月

改革开放以来，高等统计教育有了很大的发展。随着课程设置的不断调整，有不少教材出版，同时也翻译引进了一些国外优秀教材。作为培养我国统计专门人才的摇篮，中国人民大学统计学系自 1952 年创建以来，走过了风风雨雨，一直坚持着理论与应用相结合的办学方向，培养能够理论联系实际、解决实际问题的高层次人才。随着新知识和网络时代的到来，我们在教学科研的实践中，深切地感受到，无论是自然科学领域、社会科学领域的研究，还是国家宏观管理和企业生产经营管理，甚至在人们的日常生活中，信息需求量日益增多，信息处理技术更加复杂，作为信息技术支柱的统计方法，越来越广泛地应用于各个领域。

面对新的形势，我们一直在思索，课程设置、教材选择、教学方式等怎样才能使学生适应社会经济发展的客观需要。在反复酝酿、不断尝试的基础上，我们决定与统计学界的同仁，共同编写、出版一套面向 21 世纪的统计学系列教材。

这套系列教材聘请了中科院院士、中国科技大学陈希孺教授，上海财经大学数量经济研究院张尧庭教授，中国科学院数学与系统科学研究所冯士雍研究员等作为编委。他们长期任中国人民大学的兼职教授，一直关心、支持着统计学的学科建设和应用统计的发展。中国人民大学应用统计科学研究中心 2000 年已成为

国家级研究基地，这些专家是首批专职或兼职研究人员。这一开放性研究基地的运作，将有利于提升我国应用统计科学研究的水平，也必将进一步促进高等统计教育的发展。

这套教材是我们奉献给新世纪的，希望它能够为促进应用统计教育水平的提高增添一份力量。这套教材力求体现以下特点：

第一，在教材选择上，主要面向经济类统计学专业。选材既包括统计教材也包括风险管理与精算方面的教材。尽管名为统计学系列教材，但并不求大、求全，而是力求精选。对于目前已有的内容较为成熟、适合教学需要、公认的较好的教材，并未列入本次出版计划。

第二，每部教材的内容和写作，注意广泛吸收国内外优秀教材的成果。教材力求简明易懂、内容系统和实用，注重对统计方法思想的阐述，并结合大量实际数据和实例说明统计方法的特点及应用条件。

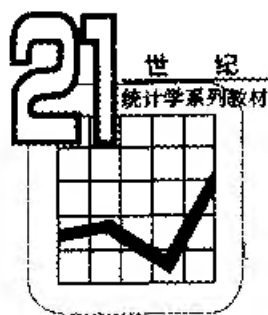
第三，强调与计算机的结合。为着力提高学生运用统计方法分析解决问题的能力，教材所涉及的统计计算，要求运用目前已有的统计软件。根据教材内容，选择使用 SAS、SPSS、TSP、STATISTICA、EViews、MINITAB、Excel 等。

感谢中国人民大学出版社的同志们，他们怀着发展我国应用统计科学的热情和提高统计教育水平的愿望，经过反复论证，使这套教材得以出版。感谢参与教材编写的同行专家、统计学系的教师。愿大家的辛勤劳动能够结出丰硕的果实。我们期待着与统计学界的同仁，共同创造应用统计辉煌的明天。

易丹辉

2000 年 8 月

于中国人民大学



前言

抽样调查是统计学专业的专业基础课,对于非统计专业的学生,有关抽样调查的知识和技能也非常重要。抽样调查在国际上已有很长的发展历史,它是政府部门、各社会团体、企业单位了解情况和搜集信息的最主要方式。近年来,抽样调查在我国得到了广泛的应用。实践证明,抽样调查是搜集信息资料的一种科学方法和手段。在信息化的今天,抽样技术在我国必将有更广泛的推广和应用。

从国际上看,虽然抽样调查的理论与方法有了很大发展,但作为讲授这门知识的基础课程而言,其内容体系已经比较成熟。综观国际间流行的有关抽样技术的教科书,其基本内容大致相同,这些基本内容主要指不同的抽样方法设计,包括简单随机抽样、分层随机抽样、整群抽样、系统抽样、多阶段抽样、比率估计、回归估计。也有一些教科书还包括了二重抽样、不等概抽样等。上述内容在本书中均有讨论。当然,能够对这些知识很好地进行论述,在有限的篇幅内把相关内容讲清、讲透,也不是一件容易的事,但我们努力去做了。

在本书的写作过程中,编著者们参阅了大量的参考文献,在汲取他人所长的同时,结合自己的教学经验和从事抽样调查项目的实践,做一些总结、归纳和概括。本书有以下特色:

1. 强调抽样技术的实际应用。抽样技术有很强的理论性,但我们仍把它看

成是一门应用性课程，在论述中侧重于方法的应用，如不同方法的应用场合、应用条件、不同方法的特点比较等。为了与全书的基调和风格一致，本书没有拘泥于理论推导，而是将必要的数学推导放在各章后的附录中。若略去这些推导，并不妨碍对书中内容的理解。本书的一部分例题和一部分习题以我们所从事过的实际调查项目为背景。习题中涉及的计算部分，均给出了参考答案，便于学习者核对。

2. 书中有两章内容在其他教科书中不多见，但却非常实用。一章是“复杂样本的方差估计”，从理论上讲这一章虽然复杂一些，但符合现代抽样技术的发展趋势，计算机技术的发展也为复杂样本的方差估计提供了方便。事实上，许多方差估计软件中的算法就是取自其中。另一章是“调查中的非抽样误差”。大量抽样调查的实践表明，非抽样误差正在成为影响调查数据质量的一个十分重要的因素。本章讨论了几种主要的非抽样误差产生的原因、非抽样误差的测定模型、控制非抽样误差的方法以及对由于无回答造成缺失数据进行调整的方法。

3. 加强案例分析。本书选取美国人口状况调查（current population survey, CPS）作为案例，用一章篇幅予以介绍和剖析。之所以选择 CPS，是因为它是国际上最著名的大型居民入户抽样调查项目之一，有 60 多年的发展历史，集世界调查统计学家思想之精华，其设计科学、巧妙，是抽样调查中的经典之作。我们从 CPS 的设计与方法中可以得到许多有益的思考与借鉴。

4. 加强抽样技术与计算机的结合。抽样调查中一项十分重要而又繁杂的工作是计算估计量方差，但目前传统的统计软件还无法直接计算不同抽样设计的估计量方差。针对这种情况，本书在附录中用一定篇幅介绍了方差估计的计算机专用软件。这部分包括两方面的内容，一个是目前国际上常用的方差估计软件的一般性介绍，另一个是对“PC CARP”软件使用的具体介绍。该软件的功能比较齐全，能够满足通常条件下的方差估计，它最主要的特点是操作比较简单，价格比较便宜，更适合于在发展中国家推广和使用。

本书可以作为统计学专业学生抽样调查方面课程的教材，也可以用作非统计专业学生和各类人员学习抽样技术的教材或学习参考书。本书涉及内容较多，学习中可以根据不同的需求，有所取舍。

本书由金勇进博士、蒋妍博士、李序颖博士共同编写。金勇进编写第 1、6、11、12 章，并负责本书编写大纲的设计、书稿的组织和全书最后的统纂；蒋妍编写第 7、9、10 章及附录 1、2；李序颖编写第 2、3、4、5、8 章。书中的部分习题选自所列的参考书目，恕不一一列举。书中的大多数例题，来自编著者所做项目的实际案例，或借鉴其他参考书中的例题进行设计，个别典型的例题数据取

自于其他书中，在引用处均有注明。在此特向有关作者和出版社表示谢意。

在本书写作过程中，得到了各方面的大力支持。编写大纲经过教材编委会的认真讨论。中国科学院数学与系统科学研究院冯士雍研究员，中国人民大学倪加勋教授对本书的初稿进行了仔细、认真的审阅，提出了许多宝贵意见。在此，特向他们表示由衷的感谢。最后我们要感谢中国人民大学出版社为出版本书给予的大力支持。

尽管我们尽了最大努力，但书中仍会有一些缺憾。对于书中的不足，恳请各位专家和读者提出宝贵意见。

金勇进

2002年3月



目 录

第1章 绪 论	(1)
§ 1.1 统计信息与抽样调查	(1)
§ 1.2 基本概念	(6)
§ 1.3 几种基本的抽样方法	(11)
§ 1.4 抽样调查程序	(13)
小 结	(16)
习题	(16)
第2章 简单随机抽样	(18)
§ 2.1 引 言	(18)
§ 2.2 估计量	(22)
§ 2.3 样本量的确定	(27)
§ 2.4 其他问题	(32)
小 结	(34)
本章附录 简单随机抽样简单估计量性质的证明	(34)
习题	(37)

第3章 分层随机抽样	(40)
§ 3.1 引 言	(40)
§ 3.2 估计量	(43)
§ 3.3 样本量在各层的分配	(47)
§ 3.4 样本量的确定	(50)
§ 3.5 分层时的若干问题	(55)
小 结	(60)
本章附录 分层抽样估计量性质的证明	(60)
习题	(63)
第4章 比率、回归与差值估计	(67)
§ 4.1 引 言	(67)
§ 4.2 比率估计	(69)
§ 4.3 回归估计	(79)
§ 4.4 差值估计	(86)
小 结	(87)
本章附录 比率估计量、回归估计量性质的证明	(88)
习题	(92)
第5章 不等概抽样	(95)
§ 5.1 引 言	(95)
§ 5.2 放回不等概抽样	(99)
§ 5.3 不放回不等概抽样	(103)
小 结	(111)
本章附录 不等概抽样估计量性质的证明	(111)
习题	(112)
第6章 整群抽样	(114)
§ 6.1 引 言	(114)
§ 6.2 群规模相等时的估计	(116)
§ 6.3 群规模不等时的估计	(122)
§ 6.4 总体比例的估计	(131)
小 结	(134)

本章附录 整群抽样估计量性质的证明	(134)
习题	(136)
第7章 系统抽样	(141)
§7.1 引言	(141)
§7.2 等概率系统抽样估计量	(147)
§7.3 不同特征总体的系统抽样	(152)
§7.4 系统抽样的方差估计	(158)
小结	(162)
本章附录 不同特征总体系统抽样的性质证明	(163)
习题	(165)
第8章 多阶段抽样	(168)
§8.1 引言	(168)
§8.2 初级单元大小相等的二阶抽样	(171)
§8.3 初级单元大小不等的二阶抽样	(177)
§8.4 其他问题	(183)
小结	(190)
本章附录 多阶段抽样估计量性质的证明	(190)
习题	(193)
第9章 二重抽样	(195)
§9.1 引言	(195)
§9.2 为分层的二重抽样	(198)
§9.3 为比率估计的二重抽样	(202)
§9.4 为回归估计的二重抽样	(206)
小结	(208)
本章附录 二重抽样公式的证明	(208)
习题	(210)
第10章 复杂样本的方差估计	(213)
§10.1 引言	(213)
§10.2 随机组方法	(216)

§ 10.3 平衡半样本方法	(226)
§ 10.4 刀切法	(236)
§ 10.5 泰勒级数法	(242)
§ 10.6 方法的比较	(244)
小 结	(245)
本章附录 复杂样本的方差估计的性质证明	(245)
习题	(247)
第 11 章 调查中的非抽样误差	(249)
§ 11.1 引 言	(249)
§ 11.2 抽样框误差	(251)
§ 11.3 无回答误差	(257)
§ 11.4 计量误差	(266)
§ 11.5 离群值的检测和处理	(272)
小 结	(275)
习题	(275)
第 12 章 设计与方法——美国 CPS 案例	(277)
§ 12.1 概 述	(277)
§ 12.2 CPS 抽样设计	(282)
§ 12.3 CPS 目标量估计	(286)
§ 12.4 CPS 的方差估计	(290)
§ 12.5 非抽样误差及控制	(296)
附录 1 方差估计软件的介绍与比较	(302)
附录 2 PC CARP 软件的基本用法	(308)
附录 3 随机数表	(338)
习题参考答案	(347)
参考文献	(353)



第 1 章

绪 论

§ 1.1 统计信息与抽样调查

一、统计信息的重要性

社会的发展离不开统计资料。对统计信息的收集和分析的实践活动很早以前就有了。我国早在2 000多年前,越国大夫范蠡(陶朱公)就曾对商品供求和价格变动之间的关系说过:“论其有余不足,则知贵贱,贵上极则反贱,贱下极则反贵。”意思是,了解市场商品供求的信息,可以预见价格的涨落,价格涨到一定限度,反而会下降;价格下降到一定限度,反而会上升。一些精明的小生产者和商人,就注意为自己的生产和经营收集市场情报,作为经营的参考。不过那时商品经济还不发达,市场规模狭小,人们对统计信息重要性的认识还远不如现代人那样深刻。

20世纪以来,生产力得到了空前发展,市场迅速扩大,企业之间的竞争日益加剧。了解市场的商情变化,了解竞争对手的情况,以此作为生产经营决策的依据,这些都需要统计信息。社会化大生产的发展,加速了自然经济的瓦解,各经济部门之间的相互依赖进一步加强。生产规模越大,就越需要以客观现实为依据,以统计信

息为依据。统计信息不仅为企业管理所需要,也为国家管理所需要。例如,政府要制定工资或价格政策,就需要居民的家庭收支、家庭生活状况和市场价格水平资料;要制定有关进出口贸易的政策,就需要各种产品的生产和使用资料;要了解人民生活的改善情况,就需要出生率、死亡率、人口平均寿命、人民受教育程度及物质和精神文化消费方面的资料;等等。现代科学技术的迅猛发展,尤其是以计算机为核心的信息处理技术的迅速发展,使得信息逐渐形成一个专门的行业部门,越来越多的人转向这个部门,从事信息的收集、处理、传递和存储等工作。人们清楚地看到,充分的信息资源和有效的信息处理技术是正确决策的必要条件,它会产生巨大的物质财富,人们称这种变化为信息时代。而统计正是获取信息的重要手段之一,统计信息是信息的重要组成部分。可以说,没有充分、准确的统计信息,就不会有科学的决策。社会越发展,对统计信息的需求也就越强烈。

二、数据的类型

统计数据展示了客观现象数量方面的特征,不同数据的性质和特点存在着差别,因此可以把统计数据分为两大类,即调查数据和试验数据。

调查数据一般是指客观上已经存在,但需要通过观察或询问才能得到的数据。例如社会现象规模、水平、相互关系和发展变化的资料。具体说调查数据有以下几个特点。首先,这类资料大多与时间有关,数据所展示的是特定时期或时点上的结果,如一定时期内的生产量、一定时点上的人口数等。其次,这类资料会随着时间的变化而改变,因此定期的收集就非常重要,因为每次收集的结果不仅展示了研究对象目前的状态,而且把以往收集的资料汇集在一起,构成时间数列,可以据此分析事物之间的相互影响和发展变化,这就为信息的进一步开发提供了广阔的空间。最后,也有一些数据,它们在短期内变化不大,相对比较稳定。最常见的就是一个国家或地区的地理和地质资料,如地形、气候条件、土壤类型、矿物储量等。这类数据的调查往往技术性强,需要这方面的专业人员使用专门的设备进行。这类调查的成本较高,而一旦取得这方面的资料就相对比较稳定,不需要经常进行。

试验数据通常与自然科学的研究相联系,其特点是在试验进行前尚未发生,因而需要通过事先的试验设计,在控制的条件下进行试验,并将试验的过程及结果加以记录和整理。试验通常是可以重复进行的,如化肥的增产效果、防治病虫害最有效的杀虫剂、某种化学变化合适的温度等。这种类型的数据往往与试验的条件有关,若改变试验的控制因素,试验结果就会发生变化。试验的次数可以是无限的。

本书所讨论的,是人文社会科学领域中的抽样调查,因此,后面所涉及的内容都是以调查数据为背景。

三、抽样调查与抽样类型

抽样调查是一种非全面性的调查,它是指从研究对象的全体(总体)中抽取一部分单位作为样本,根据对所抽取的样本进行调查,获得有关总体目标量的了解。这是广义的抽样调查的概念。

从总体抽取样本的方法看,可以分为两类抽样:一类是非概率抽样;一类是概率抽样

非概率抽样没有严格的定义,这类抽样有许多不同的具体抽取样本的方法。我国社会经济统计学教科书中谈到的重点调查和典型调查,市场调查教科书中谈到的有目的抽样、判断抽样、方便抽样和定额抽样等都属于非概率抽样。非概率抽样的共同特点是,抽取样本时不是按照随机原则,而是根据主观判断有目的、有意识地进行,或根据方便的原则进行。不同的非概率抽样方法都有各自的特点,如便于组织、节省费用、迅速快捷等等,因此不论对政府统计而言,还是对市场调查而言,非概率抽样方法都是不可缺少的。但是,采用非概率抽样方法获得的数据不能用来计算抽样误差,不能从概率的意义上控制误差并以此来保证推断的准确性。因此,如果调查的目的是用样本数据推断总体的目标量,并以一定的把握程度保证总体目标量落在目的范围,这时非概率抽样是不适合的。

概率抽样也称随机抽样,它具有以下几个特点:

1. 按一定的概率以随机原则抽取样本 所谓随机原则就是在抽取样本时排除主观上有意识地抽取调查单元,使每个单元都有一定的机会被抽中。需要注意的是,随机不等于“随便”,随机有严格的科学含义,可以用概率来描述,而“随便”则带有人为主观的因素。例如,要在一栋楼内抽取 10 位居民作为样本,若采用随机原则,就需要事先将居住在该楼的居民按某种顺序编上号,通过一定的随机化程序,如使用随机数表抽取出样本,这样可以保证居住在该楼的每位居民都有一定的机会被选中。而如果调查人员站在楼前,将最先走出楼外的 10 位居民选入样本,就是随便而不是随机,这种方法不能使每个单元都有一定的机会被选中,已经在楼外的人不可能被选中,由于某些原因在调查时段不可能外出的人也没有机会被选中。随机与随便的本质区别在于,是否按照给定的人样概率,通过一定的随机化程序抽取样本单元

2. 每个单元被抽中的概率是已知的,或是可以计算出来的。

3. 当用样本对总体目标量进行估计时,要考虑到该样本(或每个样本单元)被抽中的概率。这就是说,估计量不仅与样本单元的观测值有关,也与其人样概率有关

需要提及的是,概率抽样与等概率抽样是两个不同的概念。当我们谈到概率抽

样时,是指总体中的每个单元都有一定的非零概率被抽中,单元之间被抽中的概率可以相等,也可以不等。若是前者,称为等概率抽样;若是后者,称为不等概率抽样。

概率抽样最主要的优点是,可以依据调查结果计算抽样误差,从而得到对总体目标量进行推断的可靠程度。从另一个方面讲,也可以按照要求的精确度,计算必要的样本单元数目,所有这些,都为对调查方案的评估提供了有力的依据。

本书后面讨论的抽样调查方法,均是对概率抽样而言,因此可以把狭义的抽样调查视为概率抽样调查

四、抽样调查的作用

1. 节约费用。抽样调查能节约人力、物力和财力,从而大大降低调查费用。特别是当总体较大时,抽样调查只调查总体中的一小部分,因而节约费用的特点表现得尤为突出。

2. 时效性强。有些调查具有很强的时效性,要求在较短的时间内完成并提供调查数据。与全面调查相比,抽样调查所调查的单元少,数据采集和汇总整理的工作量较小,因而可以更快地提供调查结果。因此,对于时效性要求比较强的调查,通常采用抽样调查的方式。

3. 可以承担全面调查无法胜任的项目。有些事物或客观现象,需要通过调查掌握其数据,但又不可能进行全面调查,这时必须采用抽样调查,如居民的家庭收支状况、电视节目的收视率,以及观察或测试具有破坏性,如显像管的寿命、种子的发芽率等,这些项目的调查只能采用抽样的方法。

4. 有助于提高调查数据的质量。虽然抽样调查只调查总体中的一小部分,用部分的调查结果推断总体,存在着抽样误差,但这只是问题的一个方面。抽样调查节约费用、时效性强,在一些情况下,会比全面调查得到更准确的结果。这是因为一项调查的误差来自于多个方面,全面调查由于参与的人员多、涉及的范围大,虽然没有抽样误差,但在数据采集和数据汇总整理过程中却有产生其他误差的可能性,所以调查规模并不是越大越好。与全面调查相比,抽样调查的工作量小,这就为使用素质较高的工作人员并对其进行深入培训创造了条件。此外,可以对调查过程进行更为仔细的监督、检查和指导,使得抽样调查所得到的数据质量比同样的全面调查数据质量更高,从而使调查的总误差更小。

五、抽样调查与普查

普查是一种全面调查的方法,它是指对研究总体中的所有单元逐一进行的调

查 与全面调查相比,抽样调查虽然有许多特点和长处,但抽样调查不能代替普查,它们各有自己特殊的作用。对于有关国计民生的重要现象,有时需要了解总体中每个单元的情况,这时就需要采用普查。普查不仅能够提供研究对象的总体情况,还可以提供各个区域、各种类别的统计信息,以满足各级政府行政管理的需要,而抽样调查在这些方面则存在局限。普查资料还是构造抽样框的极好素材,所以抽样调查要与普查相结合,相互补充。它们之间这种相辅相成的关系,表现在以下几个方面

1. 抽样调查作为普查的补充。前面提到,对于有关国计民生的重要现象,需要采用普查的方法,了解总体中每个单元的基本情况,如我国进行过的普查就有人口普查、全国基本单位普查、土地资源普查等。但每一次普查都需要很大的财力投入,不可能经常进行。这时可以在两次普查之间,采用抽样调查的方法,对该种现象的变化情况进行估计。例如,现在我国每 10 年进行一次全国性的人口普查,而中间的每年进行一次人口变动量的抽样调查,对当年的人口出生、死亡、迁移等情况进行估计,抽样调查对普查起到了补充的作用。

2. 用抽样调查对全面统计资料进行评估和修正。例如,在一项普查结束后,通常采用抽样调查的方法,对随机抽取出的一部分单位进行认真仔细的复核,对发生错误的原因进行分析,计算误差率,作为对普查结果进行质量评估和数据修正的依据。

3. 利用抽样调查作深层次分析。由于普查的范围广,接受调查的单位多,因而调查项目不可能太多。在普查的基础上,根据研究的需要,可以针对某些问题,采用抽样调查的方法,获得更为详尽的资料,进行深层次的分析。

4. 利用抽样调查,提前获得总体目标量的估计。普查所涉及的单位多,数据浩繁,整理汇总工作需要较长时间,为了尽快得到总体某些特征的数据,可以采用抽样的方法,提前得到这些主要目标量的估计。

5. 普查为抽样框提供资料。普查或其他全面调查资料(例如某些统计报表)可以为抽样调查所需要的抽样框提供资料,提供辅助信息以提高抽样效率,同时也为样本轮换等提供基础资料。

六、抽样调查的应用领域

近几十年来,抽样调查的理论和实践有了迅速发展,抽样调查的应用越来越广泛。政府部门采用抽样调查的方法收集统计信息,为制定政策、进行管理提供依据;学术机构、社会团体和企业也广泛采用抽样调查的方法收集数据,进行学术研究,了解社会情况,了解市场状况。可以说,凡是需要统计数据的领域,都有可能采用抽

样调查。概括起来,抽样调查常常用于以下方面。

1. 社会经济现象的调查。社会经济涵盖的范围十分广泛。目前我国政府统计部门制度化的抽样调查项目主要有:人口变动抽样调查;农产量抽样调查;城市居民住户抽样调查;农村经济抽样调查;小型工业企业生产情况调查;小型商业企业交易情况调查;物价调查;等等。各部委根据自身业务情况进行的抽样调查项目就更多。抽样调查已成为政府部门获取统计信息的重要方式。

2. 社会性的民意调查。民意调查在西方国家十分盛行,从总统选举到居民居住小区的改造,凡是人们关心的社会问题一出现,马上就有相应的调查活动伴随。频繁的调查活动培育出一批世界闻名的,如盖洛甫那样的调查机构。过去,我国的社会性民意调查基本上是一片处女地,随着改革开放的不断推进,民意调查为越来越多的人所重视,人们开始在这片沃土上耕耘。报刊上经常见到由各种学术单位和调查机构进行民意调查的调查报告,问题选择之精妙、涉及领域之宽泛都是过去的调查所无法比拟的。社会性的民意调查已成为调查业中一道亮丽的风景线,可以预言,人们一定会在社会性民意调查这片沃土上取得更丰硕的果实。

3. 市场调查。市场调查是企业获取市场信息的主要工具。市场经济越发展,竞争越激烈,市场信息就越重要。市场调查的对象通常是消费者,通过调查,了解不同消费者群体有关商品消费的事实、动机和意向。近些年,我国的市场调查发展很快,涌现出大量的从事市场调查、咨询的专业性机构。我国人口众多,是一个巨大的消费市场,市场调查在我国有巨大的发展潜力。

§ 1.2 基本概念

一、目标总体与抽样总体

目标总体也可简称为总体,是指所要研究对象的全体,它由研究对象中所有性质相同的个体组成,组成总体的各个个体称为总体单元或单位。例如,我们要研究北京市个体商业的情况,目标总体就是北京市所有从事商业活动的个体经营单位,每个个体经营单位(或摊位)就是总体单元(单位)。目标总体的划分有时比较容易,有时就不太容易。以上面个体商业的调查为例,有些个体经营单位主要从事商品生产活动,同时兼做商品的零售,这些单位是否属于个体商业单位,就是常说的统计口径问题。在一项调查中,要对目标总体的范围做出具体规定。

抽样总体是指从中抽取样本的总体。按理,抽样总体应该与目标总体完全一致,但实践中两者不一致的情况却时常发生。仍以个体商业调查为例,目标总体是

北京市个体商业经营单位,抽样总体是什么呢?这时可以有不同的选择,选择之一是营业执照,即把北京市工商局个体商业的营业执照记录作为抽样总体,从中抽取样本。但是,有些人虽然持有营业执照,但早已不再从事商品交易活动,他们已不属于目标总体范围,但却出现在抽样总体当中;还有一些人无照经营,他们应该属于目标总体范围,却没有出现在抽样总体之中。这表明,要保证目标总体和抽样总体完全一致,不是一件容易的事情。理想的状态是,抽样总体由目标总体所决定,但在实践中,可以构造的抽样总体却有可能反过来决定调查中的目标总体。

二、抽样框与抽样单元

抽样总体的具体表现是抽样框。通常,抽样框是一份包含所有抽样单元的名单,给每一个抽样单元编上一个号码,就可以按一定的随机化程序进行抽样。对抽样框的基本要求是,抽样框中应该具有抽样单元名称和地理位置的信息,以便调查人员能够找到被选中的单元。在电话调查中,电话号码簿便是抽样框,它起到了提供抽样单元信息的作用。好的抽样框不仅与目标总体保持一致,而且还尽可能多地提供与研究的目标量有关的辅助信息,以便调查人员利用这些辅助信息搞好抽样设计,提高抽样估计的效率。

抽样单元是构成抽样框的基本要素,抽样单元可以只包含一个个体,也可以包含若干个个体,抽样单元还可以分级。在抽样单元分级情况下,总体由若干个较大规模的抽样单元组成,这些较大规模的抽样单元称为初级单元,每个初级单元中又可以包含若干个规模较小的单元,称为二级单元。用同样的方法还可以定义三级单元、四级单元等。例如,欲对北京市小学生的视力状况进行抽样调查,可以把每所小学视为初级单元,把小学校中的班级视为二级单元,把学生视为三级单元。抽取样本的顺序为先抽取学校,再抽取班级,最后抽取学生。单元可以是自然形成的,也可以是人为划分的。在一项调查中,单元分成几级不是固定不变的。在前面的例子中,如果采用抽取小学校,然后在中选的学校中直接抽取接受调查的学生而越过班级时,学校就是初级单元,学生则成为二级单元。通常把接受调查的最小一级抽样单元称为基本抽样单元。在上面的例子中,小学生是基本抽样单元。抽样单元的不同划分,是针对不同抽样方法而言的。若抽样单元只包含一个个体,并且没有分级,与之相对应的是简单随机抽样;若抽样单元中包含若干个个体,与之对应的是整群抽样;在抽样单元分级情况下,与之对应的是多阶段抽样。由于抽样单元可以分级,于是就有了与之相对应的不同级上的抽样框。抽样实践中,抽选哪一级抽样单元,有同级的抽样框即可。

三、总体指标与样本统计量

总体指标通常是调查的目标量,是我们所要研究的总体中某种特征的数量表现。总体的指标可以有很多,这些指标值是我们所关心但又是未知的,抽样调查的目的是获得对这些目标量的估计。设总体有 N 个基本单元, Y_1, Y_2, \dots, Y_N 为各基本单元的数值,根据总体指标数学处理方式的不同,可以将总体指标分为以下几种。

1. 总体总量,也称总体总和(population total)。如某地区粮食总产量,商品零售总额等。数学表达式为:

$$Y = \sum_{i=1}^N Y_i$$

2. 总体均值,也称总体平均数(population mean)。如某地区粮食平均亩产,人均储蓄存款余额等。数学表达式为:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$

3. 总体比例(proportion)。如全部产品中合格品所占比例,赞成某项政策的人所占比例等。数学表达式为:

$$P = \frac{\sum_{i=1}^N Y_i}{N}, \text{ 当第 } i \text{ 单元具有某个特定的特征时, } Y_i = 1, \text{ 否则 } Y_i = 0$$

4. 总体比率(population ratio)。它是两个总体总量或总体均值之比。如固定资产利用率,人均可支配收入变动率等。数学表达式为:

$$R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}$$

式中, Y, \bar{Y} 和 X, \bar{X} 为两个总体指标值。

把从总体中按一定程序抽出的部分总体基本单元的集合称为样本。称样本中包含基本单元的个数 n 为样本量。样本统计量是用样本中 n 个基本单元的数据构造的,作为对总体目标量的估计。统计量是样本的函数,它是随机变量,其结果取决于抽样设计和被选人样本的总体基本单元的特定组合。统计量的真正价值不在于自身的结果是多少,而在于提供有关总体目标量的信息。研究统计量的数学期望和方差是抽样理论所讨论的主要问题。

四、估计量方差、偏倚、均方误差

在抽样调查中,把样本统计量作为目标量的估计量,样本值便是目标量的估计值。样本统计量是一个随机变量,在随机原则下抽取出的不同样本,即使每个样本

的样本量 n 相同,而且根据同样的抽样设计,来自于同一个总体,它们各自的结果也会不同。估计值与总体指标值(待估参数)之间存在着离差(或差异)。这种离差有两个特点:首先,它们是不同的,有些估计值与待估参数的离差大,有些离差小;有些离差为正值,有些离差为负值。其次,这些离差虽然客观存在,但却是未知的,因为待估参数的具体值我们并不知道。抽样理论要回答抽样误差问题,因此考虑估计值与待估参数之间的差异,就只能从概率的角度去陈述,即如果相同的抽样重复多次,估计值的变化情况如何,会出现哪些结果,每个结果出现的概率是多少,离差会在什么样的范围内变化,等等。所有这些,就构成了估计量的分布,我们把估计量分布的方差称为估计量方差,它是从平均的意义上说明估计值与待估参数的差异状况,也是对抽样方案进行评价的标准之一。从这个意义上说,一个抽样设计方案比另一个抽样设计方案好,是因为它的估计量方差小。从直观上看,就是按这种方案多次抽取样本,大多数的估计值更靠近待估参数值,这意味着抽到一个好样本的可能性更大。对估计量方差开方便得到估计量标准差,也称为标准误差或标准误。它的作用与估计量方差类似。

偏倚是指按照某一抽样方案反复进行抽样,估计值的数学期望与待估参数之间的离差。设待估参数为 θ , 其估计值为 $\hat{\theta}$, 则偏倚的定义为:

$$\text{偏倚} = E(\hat{\theta}) - \theta \quad (1.1)$$

偏倚与估计量方差不同,估计量方差是由于抽样的随机性而产生的一种随机性误差,没有系统性,偏倚则是偏于某个方向的系统性误差。此外,估计量方差可以随样本量的增大而减小,而大多数的偏倚(少数有偏估计量除外)并不随样本量的增大而减小。偏倚产生的原因有两种情况,一种是估计量本身是有偏的,这时估计量的数学期望与总体参数不一致;另一种情况是非抽样误差因素的影响。

在没有偏倚的情况下,用样本统计量对目标量进行估计,要求估计量的方差越小越好。如果存在偏倚,就需要把估计量方差和偏倚综合起来加以考虑,由此提出了均方误差的概念。均方误差指所有可能的估计值与待估参数之间离差平方的均值,它等于估计量方差加偏倚的平方。令待估参数为 θ , 其估计值为 $\hat{\theta}$, 估计值的数学期望为 $E(\hat{\theta})$, 则均方误差 MSE(mean square error) 为:

$$\begin{aligned} \text{MSE} &= E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - E(\hat{\theta}) + [E(\hat{\theta}) - \theta]]^2 \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 + 2E[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta] \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 = V(\hat{\theta}) + B^2 \end{aligned} \quad (1.2)$$

式中, $V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$ 为估计量方差; $B^2 = [E(\hat{\theta}) - \theta]^2$ 为偏倚的平方。如

果估计量 $\hat{\theta}$ 的偏倚为零,也即满足

$$E(\hat{\theta}) = \theta \quad (1.3)$$

则称 $\hat{\theta}$ 为无偏估计量。对于无偏估计量,它的均方误差等于它的估计量方差。根据式(1.2),可以将估计量方差、偏倚、均方误差的关系用图 1.1 表示。

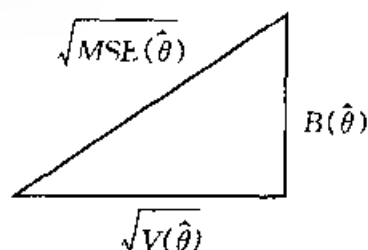


图 1.1 估计量方差、偏倚及均方误差的关系

由于偏倚是一种系统性误差,因而在抽样调查中应尽量避免。但是,也有一些估计量是有偏的,然而由于偏倚小,估计量方差也比较小,从而使均方误差比较小,这时选择这些有偏的估计量并不是一件坏事。一般说来,人们更倾向于把均方误差 MSE 作为评价抽样方案优劣的准则。

五、抽样误差与非抽样误差

抽样误差是抽取样本的随机性造成的样本值与总体值之间的差异,只要采用抽样调查,抽样误差就不可避免。抽样误差是一个一般的概念,可以用不同的量值来表示。估计量方差 $V(\hat{\theta})$ 及估计量标准差 $\sqrt{V(\hat{\theta})}$ 都是抽样误差的表现形式。在抽样调查中,抽样误差虽无法消除,但可以对其进行计量并加以控制。控制抽样误差的根本方法是改变样本量。在其他条件相同的情况下,样本量越大,抽样误差越小。抽样误差与样本量的平方根大致成反比关系,如图 1.2 所示。

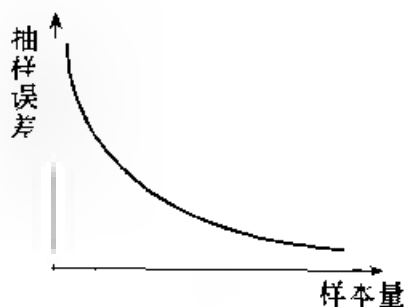


图 1.2 抽样误差与样本量的关系

由图 1.2 可以看出,抽样误差在开始时随样本量的增大而显著缩小,但经过一定阶段后便趋于稳定。也就是说,经过一定阶段后,用增大样本量的方式减少抽样误差一般是不合算的。这时,只要稍微降低一些精度,就可以大幅度减少样本量从而节省可观的调查费用

非抽样误差是相对于抽样误差而言的,它不是由于抽样的随机性,而是由于其他多种原因引起的估计值与总体参数之间的差异。例如,由于调查计划不周、调查对象范围划分不清而产生的误差;构造抽样框时,目标总体与抽样总体不一致所带来的抽样框误差;调查过程中由于无回答或回答有误造成的误差;填写调查表以及数据录入和汇总过程中产生的误差;等等。非抽样误差问题将有专门讨论。

六、精度与费用

通常,精度由误差来表现。如果不考虑非抽样误差,则精度的具体体现就是抽样误差。抽样误差越小,说明用样本统计量对总体参数进行估计时的精度越高。抽样误差与样本量有关,样本量越大,在其他条件相同情况下,抽样误差就越小,抽样调查的精度就越高。同时,样本量也与调查费用有关,样本量越大,调查费用就越高。样本量与调查费用大致呈线性关系,但样本量与精度却呈非线性关系。也就是说,在样本量比较小时,每增加一个抽样单元对提高精度的作用比较显著,但随着样本量的增大,达到一定程度后,再每增加一个抽样单元对提高精度的作用就逐渐下降。

除了样本量以外,影响精度与费用的另外因素是抽样方式与估计方法。一个好的抽样设计必须同时考虑到精度与费用两个方面。反过来,精度与费用也是评价抽样设计方案优劣的两条准则。对于一个具体的抽样设计,在核定的费用内达到最高的精度,或在达到精度要求的条件下使调查的费用最少,则称这样的抽样设计为最优设计。最优设计的抽样效率最高,因此效率是对精度与费用的综合。

§ 1.3 几种基本的抽样方法

一、简单随机抽样(simple random sampling)

简单随机抽样也称纯随机抽样,是从抽样框内的 N 个抽样单元中随机地、一个一个地抽取 n 个单元作为样本,在每次抽选中,所有未入样的待选单元入选样本的概率是相等的,这 n 个被抽中的单元就构成了简单随机样本,简单随机样本也可

以一次同时从总体(抽样框)中抽出,这时全部可能样本中的每一个样本被抽中的概率也需要相等。抽样的随机性是通过抽样的随机化程序体现的,实施随机化程序可以使用随机数字表,也可以使用能产生符合要求的随机数序列的计算机程序。

简单随机抽样是一种最基本的抽样方法,是其他抽样方法的基础。这种方法的突出特点是简单直观,在抽样框完整时,可以直接从中抽选样本,由于抽选的概率相同,用样本统计量对目标量进行估计及计算抽样误差都比较方便。但简单随机抽样在实际应用中也有一些局限,首先,它要求包含所有总体单元的名单作为抽样框,当 N 很大时,构造这样的抽样框并不容易。其次,根据这种方法抽出的单元很分散,给实施调查增加了困难。最后,这种方法没有利用其他辅助信息以提高估计的效率。所以在规模较大的调查中,很少直接采用简单随机抽样,一般是将这种方法同其他抽样方法结合在一起使用。

二、分层抽样(stratified sampling)

分层抽样是将抽样单元按某种特征或某种规则划分为不同的层,然后从不同的层中独立、随机地抽取样本。将各层的样本结合起来,对总体的目标量进行估计。分层抽样有许多优点,它保证了样本中包含有各种特征的抽样单元,样本的结构与总体的结构比较相近,从而可以有效地提高估计的精度;分层抽样在一定条件下为组织实施调查提供了方便,如果层的划分是按行业或行政区划进行的就是这样;分层抽样既可以对总体参数进行估计,也可以对各层的目标量进行估计;等等。这些优点使分层抽样在实践中得到了广泛的应用。

三、整群抽样(cluster sampling)

将总体中若干个基本单元合并为组,这样的组称为群。抽样时直接抽取群,然后对中选群中的所有基本单元全部实施调查,这样的抽样方法称为整群抽样。

与简单随机抽样相比,整群抽样的优点在于,首先,抽取样本时只需要群的抽样框而不必要求具有所有基本单元的抽样框,这就大大简化了编制抽样框的工作量。其次,由于群通常是由那些地理位置邻近、或隶属于同一系统的单元所构成,因此调查的地点相对集中,从而节省了调查费用,便于调查的实施。整群抽样的主要缺点是估计的精度较差,因为同一群内的单元或多或少地有些相似,在样本量相同条件下,整群抽样的抽样误差通常比较大。一般说来,要得到与简单随机抽样相同的精度,采用整群抽样需要增加基本调查单元。

四、系统抽样(systematic sampling)

将总体中的所有单元(抽样单元)按一定顺序排列,在规定的范围内随机地抽取一个单元作为初始单元,然后按事先规定好的规则确定其他样本单元,这种抽样方法称为系统抽样。典型的系统抽样是先从数字1到 k 之间随机抽取一个数字 r 作为初始单元,以后依次取 $r+k, r+2k, \dots$ 单元,所以可以把系统抽样看成是将总体内的单元按顺序分成 k 群,用相同的概率抽出一群的方法。

系统抽样的主要优点是操作简便,如果有辅助信息,对总体内的单元进行有组织的排列,可以有效地提高估计的精度。系统抽样的缺点是对估计量方差的估计比较困难。系统抽样方法在调查实践中有广泛的应用。

五、多阶段抽样(multi-stage sampling)

采用类似整群抽样的方法,首先抽取群,但并不是调查群内的所有基本单元,而是再进行一步抽样,从选中的群中抽取出若干个基本单元进行调查。因为取得这些接受调查的基本单元需要两个步骤,所以将这种抽样方法称为二阶段抽样。这里,群是初级抽样单元,第二阶段抽取的是基本抽样单元。将这种方法推广,使抽样的段数增多,就称为多阶段抽样。例如第一阶段抽取初级单元,第二阶段抽取二级单元,第三阶段抽取接受调查的基本单元就是三阶段抽样,用同样的方法还可以定义四阶段抽样。不过,即便是大规模的抽样调查,抽取样本的阶段也应当尽可能地减少。因为每增加一个抽样阶段,就会增加一份抽样误差,用样本对总体进行估计也更加复杂。

多阶段抽样具有整群抽样的优点,它保证了样本相对集中,从而节约了调查费用;不需要包含所有低阶段抽样单元的抽样框;由于实行了再抽样,使调查单元在更大的范围内展开。在较大规模的抽样调查中,多阶段抽样是经常被采用的方法。

§ 1.4 抽样调查程序

对于不同的抽样调查项目,整个调查过程所包含的步骤不尽相同。但一般而言,都需要以下几个步骤。

一、确定调研问题

这是整个调查的第一步,也是至关重要的一步。在这个过程中首先需要明确地定义问题,包括对整个问题的叙述以及确定研究问题的具体组成部分。只有问题定义清楚了,才有可能进一步设计和执行。确定调研问题所要回答的是“要做什么样的调查研究”和“为什么要做这项调查研究”。调研人员需要考虑研究的目的、相关的背景材料、所需要的信息以及这些信息在进行分析时如何使用。为此,调研人员需要与有关部门的决策者进行认真讨论,访问有关行业的专家,分析二手资料,必要时还需要进行如座谈会那样的定性调查。

在这一过程中,还要考虑到财力限制及有关的调查技术力量,把调研的问题定义在适当的范围内。每一项调查,都会有费用、时间等方面的要求。比如一项大规模的调查,需要较多的调查费用,而实际的预算费用明显不够,就必须缩小调研问题的范围以适应财力的许可。

二、抽样方案设计

抽样方案要描述样本是如何抽取的。调查中有不同的数据收集方法,如面访调查、电话调查、邮寄调查等。不同收集方法需要不同的抽样框,抽样方案设计包括抽样框的设计。此外,对样本又有不同的抽取方法,在制定抽样方案时,既要考虑方法的科学性,又要照顾实际的可行性。例如,在—项多阶段抽样中,前一二阶抽样十分关键,需要采用效率高的抽样方法,由于这个阶段的抽样可以由设计人员来实施,所以技术复杂一些也无妨,后面阶段的抽样则力求简单,以便基层的操作者能够胜任。在这个过程中还要确定样本量,要给出与抽样设计相匹配的总体参数的估计公式及估计量的精度公式。调查中常常会遇到调查对象失访,如受访者不在家或拒访,因此需要制定一些具体的处理办法,把失访对调查结果的影响降到最小程度。

三、问卷设计

问卷设计是根据调查目的和要求,将比较抽象的调研问题逐步细化,演变为现场调查中向受访者询问的、比较具体的问题这样一个工作过程。问卷设计也是一门技巧性很强的学问。一份设计精巧的问卷,应当使受访者能准确无误地理解调查的内容,能够正确回答并且愿意回答所提的问题,并且使调查机构便于对问卷进行计算机处理,有效地利用调查数据进行统计分析。进行问卷设计,除了应具备所涉及调查内容的专业知识外,还需要有统计学、社会学、心理学及计算机等多方面的知

识,此外还要有问卷设计的技巧和经验。通常,设计出的调查问卷的初稿,应由有关方面的人士和专家进行审阅和讨论

四、实施调查过程

在这个过程中要获得样本单元的调查数据,关键的问题是要保证原始数据的质量,这就需要对调查过程进行有效的管理和监控。调查实施前,需要对调查员进行技术培训,使调查员熟悉调查问卷,掌握访谈技巧,并增强责任心。在调查过程中加强质量检验,出现问题及时总结,及时补救。调查人员要有操作手册,调查过程中也要有管理制度和措施,使得从事具体调查的人员有章可循。如果调查项目比较大,又是第一次进行,或者对问卷设计的把握不够大,在正式调查实施前,还应当进行一次预调查(试验调查),以检验各方面的工作是否完善。

五、数据处理分析

数据处理分析是调查的收获阶段,它为撰写调查报告提供基本的素材。在这个阶段,首先要对经过调查获得的原始数据进行检查、核对,对验收合格的调查问卷进行编码和录入。数据录入后,多数情况下需要进行数据的预处理,为统计分析做好准备。数据的预处理包括:录入数据的再编码,它是对原编码的补充和调整,满足某些统计分析软件对编码的特殊要求,也是根据研究要求对数据的重新归类分组;对缺失值进行插补,以构造出完整的数据集;进行变量的转换,进而进行常规的统计分析;计算目标量的估计值、方差及变异系数的估计值等。必要时还需要结合研究目的进行深入的统计处理与分析。

六、撰写调查报告

调查报告可以有不同的类型。从内容上可以分为描述性报告和探索性报告;从技术角度可以分为一般报告和技术报告;从性质上可以分为普通调查报告和学术研究报告等。这里引用联合国关于抽样调查结果的一般性报告所应包含的主要项目(United Nations, 1949),内容如下:

- (1) 主题。清楚地指出此调查的目的,并提出对调查结果的使用方式。
- (2) 范围。正确地描述调查范围,包括指定的研究定义及调查的地理区域。
- (3) 调查对象。详细叙述此调查所收集的资料项目及未列表项目的原因。

(4) 资料收集方法。清楚地叙述所采用的收集资料的方法。此外,收集资料过程中遇到的所有困难及解决的方法,均应详细说明。

(5) 调查期、参考期和报告期 调查报告中必须指出调查期、参考期和报告期等经过时间。

(6) 抽样设计和估计程序 清楚说明调查中所使用的抽样单元、抽样框、样本大小和抽样方法,清楚地叙述估计时所用的公式。

(7) 结论的描述。列表资料应该以清楚且易于理解的方式列出。合适的一览表、图例或图解能使调查结果获得更快、更清楚的理解。

(8) 精确度。调查结果中应该列出估计所达到的精确程度、检验及比较的结果、对调查质量的评估。此外,还必须指出无回答者的种类、比例以及对最后结果的影响程度。

(9) 责任。主办机构及指挥此调查的机构须在报告中提出。

(10) 参考文献。须列出已发表的相关论文和报告作为参考资料。

小 结

本章分四节。第一节是对抽样调查概念、意义、作用的介绍。从抽选样本的方法看,可以分为概率抽样和非概率抽样,二者具有不同的特点,本书的内容主要是对概率抽样而言。第二节介绍了概率抽样中所涉及的一些基本概念,这些概念的运用将贯穿全书。第三节介绍了几种基本的概率抽样方法,在后面各章中将对这些方法做一详细介绍。第四节介绍了抽样调查的一般步骤。本章的目的是使读者在学习具体的抽样技术之前,对有关的问题有一个大概的了解。

习 题

1. 举例说明什么情况下适合采用非概率抽样,什么情况下适合采用概率抽样。

2. 讨论以下情况是否属于概率抽样,并说明理由:

(1) 从实验室中一个装有 100 只兔子的大笼子里抓 10 只兔子做试验,不经任何有意识的选取,抓到哪只算哪只,抓满 10 只为止。

(2) 将笼子中的 100 只兔子编上 1 ~ 100 号,任意列出 10 个数字,相应号码的兔子作为试验用的兔子。

(3) 从在场的人的钱包中随便抽出纸币,凡兔子号码尾数与纸币号码尾数相

问者即作为抽中的样本。

3. 现在利用网络进行调查的项目很多,举例说明哪些类型的调查属于概率抽样,哪些类型的调查不属于概率抽样

4. 请指出下面一些内容的调查可以采用什么材料构造抽样框(将有效性和可能性结合起来):

(1) 对北京市居民癌症病人的调查。

(2) 对北京市小学生零花钱情况的调查。

(3) 对某地区收视率情况的调查。

5. 你认为抽样调查中的哪些环节最关键,并说明理由。



第 2 章

简单随机抽样

简单随机抽样是所有概率抽样方法的基础,我们将要学习的各种抽样方法都是在其基础上发展起来的。

本章共分四节,第一节介绍简单随机抽样的定义及其抽选方法;第二节介绍估计量及其性质;第三节介绍样本量的确定原则;第四节介绍与简单随机抽样相关的若干问题。

§ 2.1 引 言

一、定义与符号

(一) 定义

简单随机抽样也称单纯随机抽样 从含有 N 个单元的总体中抽取 n 个单元组成样本,如果抽样是不放回的,则所有可能的样本有 C_N^n 个,若每个样本被抽中的概率相同,都为 $\frac{1}{C_N^n}$,这种抽样方法就是不放回的简单随机抽样。具体抽样时,通常是逐个等概率抽取样本单元,直到抽满 n 个单元为止。

简单随机抽样根据抽样单元是否放回可分为放回简单随机抽样和不放回简单随机抽样。

1. 放回简单随机抽样 当从总体 N 个抽样单元中抽取 n 个抽样单元时,如果依次抽取单元,不管以前是否被抽中过,每次都从 N 个抽样单元中随机抽取,这时,所有可能的样本为 N^n 个(考虑样本单元的顺序),每个样本被抽中的概率为 $\frac{1}{N^n}$,这种方式就是放回简单随机抽样

应当注意的是,放回简单随机抽样在每次抽取样本单元时,都将前一次抽取的样本单元放回总体,因此,总体的结构不变,抽样是相互独立进行的,这是它与不放回简单随机抽样的主要不同之处。这一点使它的数学处理相对简单。

【例 2.1】 设总体有 5 个单元(1,2,3,4,5),按放回简单随机抽样的方式抽取 2 个单元,则所有可能的样本为 $5^2 = 25$ 个(考虑样本单元的顺序),如表 2.1。

表 2.1 放回简单随机抽样所有可能的样本

1,1	2,1	3,1	4,1	5,1
1,2	2,2	3,2	4,2	5,2
1,3	2,3	3,3	4,3	5,3
1,4	2,4	3,4	4,4	5,4
1,5	2,5	3,5	4,5	5,5

2. 不放回简单随机抽样 从总体个抽样单元中依次抽取,直到抽满 n 个抽样单元,每个被抽中的单元不再放回总体,每次抽样是从总体剩下的单元中进行。

【例 2.2】 设总体有 5 个单元(1,2,3,4,5),按不放回简单随机抽样的方式抽取 2 个单元,则所有可能的样本为 $C_5^2 = 10$ 个,如表 2.2

表 2.2 不放回简单随机抽样所有可能的样本

1,2	2,3	3,4	4,5
1,3	2,4	3,5	
1,4	2,5		
1,5			

不放回简单随机抽样的样本量要受总体大小的限制,即 n 不能超过 N ,最多等于 N ,如果样本量接近或等于总体大小时,则调查几乎是或就是普查。

在实际工作中,更多地采用不放回简单随机抽样,所以以下讨论的简单随机抽样除非特别申明,都指不放回简单随机抽样

(二) 符号

在抽样调查中,人们通常用大写符号表示总体单元的标志值,用小写符号表示样本单元的标志值。总体中 N 个单元的标志值为 Y_1, Y_2, \dots, Y_N , 样本中 n 个单元的标志值为 y_1, y_2, \dots, y_n 。

调查的目的是了解总体某个标志的性质,我们称之为总体目标量(或总体指标),主要有:总体总量 Y 、总体均值 \bar{Y} 、总体中具有某种特征的单元数占总体的比例 P 、两个总体总量或两个总体均值的比率 R 等指标。

在对估计精度进行计算或推算时,要涉及到总体方差、样本方差等指标。如表 2.3

表 2.3

总 体		样 本	
Y	$\sum_{i=1}^N Y_i = Y_1 + Y_2 + \dots + Y_N$	$\sum_{i=1}^n y_i = y_1 + y_2 + \dots + y_n$	
\bar{Y}	$\frac{1}{N} \sum_{i=1}^N Y_i = \frac{Y_1 + Y_2 + \dots + Y_N}{N}$	$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{y_1 + y_2 + \dots + y_n}{n}$	
P	$\frac{A}{N} = \frac{1}{N} \sum_{i=1}^N Y_i \quad (Y_i = 0 \text{ 或 } 1)$	$p = \frac{a}{n} = \frac{1}{n} \sum_{i=1}^n y_i \quad (y_i = 0 \text{ 或 } 1)$	
R	$\frac{\sum_{i=1}^N Y_i}{\sum_{i=1}^N X_i} = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}$	$\hat{R} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{\bar{y}}{\bar{x}}$	
S^2	$\frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{N}{N-1} \sigma^2$	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$	

总体指标值上面带符号“ \wedge ”的表示由样本得到的总体指标的估计,如 $\hat{Y}, \hat{\bar{Y}}, \hat{P}, \hat{R}$ 等,称为 Y, \bar{Y}, P, R 的估计。

估计量的方差用 V 表示,如 $V(\hat{Y})$;标准差用 S 表示,如 $S(\hat{Y}) = \sqrt{V(\hat{Y})}$,而对 $V(\hat{Y})$ 的样本估计,为避免符号的累赘,不用 $\hat{V}(\hat{Y})$,而用 $v(\hat{Y})$ 表示。类似地, $S(\hat{Y})$ 的样本估计用 $s(\hat{Y}) = \sqrt{v(\hat{Y})}$ 表示。

称 $\frac{n}{N}$ 为抽样比,记为 f 。

二、抽选方法

要产生简单随机样本,首先将总体 N 个单元从 1 到 N 编号,每个单元对应一个号,如果抽到某个号,则对应的那个单元入样。要选出 n 个单元入样,通常有两种做法:抽签法和随机数法

(一) 抽签法

当总体不大时,可以用均匀同质的材料制作 N 个签,将它们充分混合,然后一次抽取 n 个签,或一次抽取一个签但不放回,接着抽下一个签直到第 n 个签为止,则这 n 个签上所示的号码表示入样的单元号。

(二) 随机数法

当总体较大时,抽签法实施起来很困难,这时可以利用随机数表、随机数骰子、计算机产生的伪随机数进行抽样。

1. 随机数表。随机数表是由数字 0, 1, \dots , 9 组成的表,每个数字都有同样的机会被抽中。用随机数表抽取简单随机样本时,可用下面几种方法。

方法一:根据总体大小 N 的位数决定在随机数表中随机抽取几列,如 $N = 678$,要抽取 $n = 5$ 的样本,则在随机数表中随机抽取 3 列,顺序往下,选出头 5 个 001 ~ 678 之间互不相同的数,如果这 3 列随机数字不够,可另选其他 3 列继续,直到抽满 n 个单元为止。

方法二:若 N 的第一位数字小于 5,且 n 不小,则方法一可能花费较多的时间。如 $N = 327$,按方法一则 000 和 328 ~ 999 的数都没有用。这时采用下面的方法可能更好,在随机数表中随机抽取 3 列,顺序往下,如果得到的随机数在 401 ~ 800 之间,则这个数字减去 400,由此 000,大于 800 以及余数大于 327 的数字被扔掉。显然这种方法比上一种方法效率高。

方法三:若 N 的第一位数字小于 5,如 $N = 327$,且 n 不小,除了按方法二产生随机数以外,还可按下面的方法产生随机数。在随机数表中随机抽取 3 列,顺序往下,如果得到的随机数大于 327,且小于 982(因为 $327 \times 3 = 981$,而 $327 \times 4 = 1308$,因此 000 及 982 ~ 999 的数字应扔掉),则用这个数字除以 327,得到的余数入样,显然这种方法也比方法一效率高。

在使用随机数表时,为克服可能的个人习惯,增加随机性,使用随机数表的页号及起始点应该用随机数产生,如随意翻开一页,闭上眼睛,将火柴随意扔到页面上,将火柴头所指的数字作为页号,用同样的方法也可以产生起始行号和起始列号。

2. 随机数骰子。随机数骰子是由均匀材料制成的正 20 面体,面上标有 0 ~ 9

的数字各 2 个。我国“运筹”牌随机数骰子一盒有 6 个不同颜色的骰子,使用时,根据总体大小 N 的位数,如 $N = 327$ 的位数为 3,则将 3 个不同颜色的骰子放入盒中,并规定每种颜色所代表的位数,如红色代表个位数,蓝色代表十位数,黄色代表百位数等,盖上盒盖,摇动盒子,使骰子充分旋转,然后打开盒盖,读出骰子所表示的数字,重复上述步骤,直到产生 n 个不同的随机数。

3. 摇奖机 各类彩票中奖号码的产生通常是由摇奖机完成的,这个过程可以从电视节目中看到。将标有数字 0 ~ 9 的 10 个球放入摇奖机中,充分搅拌,使球充分转动,直到摇出其中的一个球,记录该球所标明的数字,产生了随机数的个位数;将球放回到摇奖机中,重复上述步骤,直到摇出一个球,记录该球所标明的数字为随机数的十位数;同理产生百位数等,如此产生一个随机数。重复上述步骤,直到产生 n 个不同的随机数。

4. 计算机产生的伪随机数。不少统计软件都有现成的产生随机数的程序,使用者也可利用同余法自编产生随机数的小程序。利用计算机产生的随机数具有快捷、方便的特点,但需要注意的是,利用计算机产生的随机数是伪随机数,并不能保证其随机性,通常产生的伪随机数有循环周期,当然,我们希望产生的伪随机数循环周期越长越好。在可能的情况下,建议还是利用随机数表和随机数骰子来产生随机数。

§ 2.2 估计量

总体目标量通常有总体的总量、均值、比例、比率等指标,本章主要考虑前三种目标量。如北京市全市职工年收入、北京市职工年平均收入、北京市男性职工的比例等。

一、总体均值的估计

(一) 简单估计量的定义

在没有其他总体信息的条件下,对总体均值 \bar{Y} 的简单估计为:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (2.1)$$

即以样本均值作为总体均值的估计。

(二) 简单估计量的性质

性质 1 对于简单随机抽样, \bar{y} 是 \bar{Y} 的无偏估计。即

$$E(y) = Y \quad (2.2)$$

为表达 y 的方差,我们先定义总体的方差。通常将有限总体的方差定义为:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - Y)^2 \quad (2.3)$$

这里,我们用

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - Y)^2 \quad (2.4)$$

来表示总体方差,这种表示可以在大多数情况下使公式的表达更为简捷。

性质 2 对于简单随机抽样, y 的方差为:

$$V(\bar{y}) = \frac{N-n}{Nn} S^2 = \frac{1-f}{n} S^2 \quad (2.5)$$

式中, n 为样本量; $f = \frac{n}{N}$ 为抽样比; $1-f$ 为有限总体校正系数。

性质 3 $V(y)$ 的无偏估计为:

$$v(y) = \frac{1-f}{n} s^2 \quad (2.6)$$

式中, s^2 为样本方差。

估计量的方差 $V(y)$ 是衡量估计量精度的度量。从式(2.5)可以看出,影响估计量方差的因素有样本量 n 、总体方差 S^2 和抽样比 f 。在需要进行抽样调查的问题中, N 通常很大(如果 N 不大,则没有必要进行抽样调查,直接进行普查更好),当 $f < 0.05$ 时,可将 $1-f$ 近似取为 1,这时主要是样本量 n 和总体方差 S^2 影响估计量方差。样本量 n 越大,估计量的方差越小。当样本量一定时,总体方差 S^2 越大,估计量的方差越大。由于总体方差 S^2 是我们无法改变的,因此,在简单随机抽样的条件下,要提高估计量的精度就只有通过加大样本量来实现。

【例 2.3】 我们从某个 $N = 100$ 的总体中抽出一个大小为 $n = 10$ 的简单随机样本,要估计总体平均水平并给出置信度为 95% 的区间估计。如表 2.4。

表 2.4 简单随机样本的指标值

序号 i	1	2	3	4	5	6	7	8	9	10
y_i	4	5	2	0	4	6	6	15	0	8

解:依题意, $N = 100, n = 10, f = \frac{10}{100} = 0.1$

计算样本均值及样本方差为:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{50}{10} = 5$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{172}{9} \approx 19.1111$$

因此,对总体平均水平的估计为:

$$\hat{Y} = \bar{y} = 5$$

对 y 的方差及标准差的估计为:

$$v(\hat{Y}) = \frac{1}{n} f s^2 = \frac{1}{10} \times 0.1 \times 19.1111 \approx 1.72$$

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} \approx 1.3115$$

由置信度 95% 对应的 $t = 1.96$, 因此, 可以以 95% 的把握说总体平均水平大约在 $5 + 1.96 \times 1.3115$ 之间, 即 $2.4295 \sim 7.5705$ 之间。

注意, 本例只是为了说明计算过程, 实际工作中, 如果总体不大, 或抽样比接近于 1 时, 人们通常采用全面调查方式, 而不是采用抽样调查。

(三) 放回简单随机抽样简单估计量

对于放回简单随机抽样, 对总体均值 Y 的简单估计为:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

\bar{y} 是 \bar{Y} 的无偏估计, 其方差为:

$$V(\bar{y}) = \frac{N-1}{N} \frac{1}{n} S^2 = \frac{1}{n} \sigma^2$$

$V(\bar{y})$ 的无偏估计为:

$$v(\bar{y}) = \frac{1}{n} s^2$$

比较放回与不放回简单随机抽样简单估计量的方差公式, 注意到不放回时的方差为放回时方差的约 $1-f$ 倍, 而 $1-f < 1$, 因此不放回抽样的估计精度比放回抽样的估计精度高。

二、总体总量的估计

总体总量 (Y) 与总体均值 (\bar{Y}) 只差一个常数, 即

$$Y = N\bar{Y} = \sum_{i=1}^N Y_i$$

因此, 对总体均值的估计结果, 可以很容易地推出对总体总量的估计。

(一) 简单估计量的定义

在没有其他总体信息的条件下, 对总体总量 Y 的简单估计为:

$$\hat{Y} = Ny = \frac{N}{n} \sum_{i=1}^n y_i \quad (2.7)$$

(二) 简单估计量的性质

性质 4 对于简单随机抽样, \hat{Y} 是 Y 的无偏估计, 即

$$E(\hat{Y}) = Y \quad (2.8)$$

\hat{Y} 的方差为:

$$V(\hat{Y}) = N^2 V(y) = \frac{N^2(1-f)}{n} S^2 \quad (2.9)$$

$V(\hat{Y})$ 的样本无偏估计为:

$$v(\hat{Y}) = N^2 v(y) = \frac{N^2(1-f)}{n} s^2 \quad (2.10)$$

【例 2.4】(续例 2.3) 估计总体总量, 并给出在置信度 95% 的条件下, 估计的相对误差。

解: 依题意, $N = 100$, 由例 2.3 的计算已知:

$$y = 5, \quad s^2 \approx 19.1111$$

因此, 对总体总量的估计为:

$$\hat{Y} = 100 \times 5 = 500$$

对 \hat{Y} 方差及标准差的样本估计为:

$$v(\hat{Y}) \approx 100^2 \times \frac{1-0.1}{10} \times 19.1111 = 17200$$

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} \approx 131.1488$$

因此, 在置信度 95% 的条件下(对应的 $t = 1.96$), \hat{Y} 的相对误差为:

$$t \frac{s(\hat{Y})}{\hat{Y}} = 1.96 \times \frac{131.1488}{500} \approx 0.5141 = 51.41\%$$

三、总体比例的估计

有时调查研究的是某一类特征的单元占总体单元数中的比例(P), 如男职工人数占总职工人数的比例。这时, 将总体单元按是否具有这种特征划分为两类, 设总体中有 A 个单元具有这个特征, 如果对每个单元都定义指标值

$$Y_i = \begin{cases} 1, & \text{第 } i \text{ 个单元具有所考虑的特征} \\ 0, & \text{其他} \end{cases}, i = 1, 2, \dots, N$$

则有

$$P = \frac{A}{N} = \frac{1}{N} \sum_{i=1}^N Y_i = \bar{Y} \quad (2.11)$$

因此,总体比例的估计是总体均值估计的另一种表现形式。

(一) 估计量的定义

对于样本量为 n 的简单随机样本,如果有 a 个单元具有所研究的特征,则对总体比例 P 的估计为样本比例,即

$$p = \frac{a}{n} \quad (2.12)$$

(二) 估计量的性质

性质 5 对于简单随机抽样, p 是 P 的无偏估计。 p 的方差为:

$$V(p) = \frac{PQ}{n} \frac{N-n}{N-1} \quad (2.13)$$

式中, $Q = 1 - P = \frac{N-A}{N}$ 。

$V(p)$ 的样本无偏估计为:

$$v(p) = \frac{1-f}{n-1} pq \quad (2.14)$$

式中, $q = 1 - p$ 。

【例 2.5】 某超市开张一段时间之后,为改进销售服务环境,欲调查附近几个小区居民到该超市购物的满意度。该超市与附近几个小区的居委会取得联系,在总体中按简单随机抽样抽取了一个大小为 $n = 200$ 人的样本。调查发现对该超市购物环境表示满意或基本满意的居民有 130 位,要估计对该超市购物环境持肯定态度居民的比例,并在置信度 95% 条件下,给出估计的绝对误差和置信区间。假定这时的抽样比可以忽略。

解:已知 $n = 200, a = 130, 1 - f \approx 1$

$$p = \frac{a}{n} = \frac{130}{200} = 65\%$$

$$v(p) = \frac{1-f}{n-1} pq \approx \frac{1}{200-1} \times 0.65 \times 0.35 \approx 0.001143$$

$$s(p) = \sqrt{v(p)} \approx 0.0338$$

所以,对该超市购物环境持肯定态度的居民的比例为 65%。

在置信度 95% 条件下,估计的绝对误差为:

$$t \times s(p) = 1.96 \times 0.0338 \approx 0.0663 = 6.63\%$$

p 的 95% 置信区间为:

$$0.65 + 1.96 \times 0.0338$$

或者说,可以以 95% 的把握认为对该超市购物环境持肯定态度的居民的比例大约在 58.37% ~ 71.63% 之间。看来,该超市的购物环境还需要改善。

§ 2.3 样本量的确定

一、有关问题

(一) 费用函数

样本量的确定在抽样调查中是一个十分重要又比较复杂的问题,它受对调查精度的要求以及调查费用的限制。在简单随机抽样情况下,设调查费用函数为:

$$C = c_0 + c_1 n$$

式中, C 为总费用; c_0 为固定费用,如管理人员开支、办公费、组织、宣传、场租费等,这些费用都与样本量 n 无关; c_1 为与样本量有关的可变费用,即每调查一个样本单元所需的费用,如调查费、差旅费、礼品费等。

作为抽样方案的设计者,应该权衡精度与费用之间的关系,使调查既满足精度的要求,又节省费用。在实际工作中,通常是在总费用一定的条件下使精度最高,或在要求精度一定的条件下,使总费用达到最小。

(二) 误差限

如果只考虑调查精度对样本量的要求,则可以按统计意义对样本量进行定量的计算。对精度的要求通常以允许绝对误差(绝对误差限) d 或允许相对误差(相对误差限) r 来表示,误差限是在一定的概率保证意义下绝对或相对误差,即对参数 θ (如总体均值) 及它的估计 $\hat{\theta}$ (如样本均值),以绝对误差限表示,有

$$P(|\hat{\theta} - \theta| \leq d) = 1 - \alpha$$

或以相对误差限表示,有

$$P\left\{\frac{|\hat{\theta} - \theta|}{\theta} \leq r\right\} = 1 - \alpha$$

由于我们对总体未做任何假定,因此 $\hat{\theta}$ 的精确分布很难求得,但当样本量足够大时,可以用正态分布近似,这时绝对误差限

$$d = t \sqrt{V(\hat{\theta})} = tS(\hat{\theta}) \quad (2.15)$$

式中, t 为标准正态分布的双侧 α 分位数。如 $1 - \alpha = 90\%$, 对应的 $t = 1.645$;

1. $\alpha = 95\%$, 对应的 $t = 1.96$ 等。而相对误差限

$$r = t \frac{\sqrt{V(\hat{\theta})}}{\hat{\theta}} = t \frac{S(\hat{\theta})}{\hat{\theta}} = tC_v(\hat{\theta}) \quad (2.16)$$

式中, $C_v(\hat{\theta})$ 为 $\hat{\theta}$ 的变异系数。在实际问题中, 当总体参数 θ 未知时, 可以用其估计量 $\hat{\theta}$ 替代。于是又可以将 r 写为:

$$r = t \frac{\sqrt{V(\hat{\theta})}}{\hat{\theta}} = t \frac{S(\hat{\theta})}{\hat{\theta}} = tC_v(\hat{\theta})$$

由于 $S(\hat{\theta})$ 是样本量的函数, 因此根据对 d 或 r 的要求, 以及 $1 - \alpha$ 所对应的 t 可推算出所需要的样本量。

(三) 其他考虑因素

确定样本量除了通过定量的方法之外, 还要考虑其他一些因素。

1. 问题的重要性。对于决策比较重要的问题, 所需的信息应该比较准确, 因此样本量要大一些。

2. 所研究问题目标量的个数。如果所研究的问题目标量较多, 样本量应适当放大。

3. 参照同类调查。参照以往同类型调查项目确定样本量。

4. 调查表的回收率。调查过程中, 可能有些调查对象拒访或因种种原因调查不到, 这时样本量应适当放大。一种做法是, 根据估计的回收率反算出应接触的样本量, 例如回收率估计为 80% , 则应接触的样本量为计算出所需样本量的 1.25 倍。

5. 有效样本。调查过程中, 可能有些接触的对象不是“合格”对象, 我们称“合格”对象为有效样本。为了获得足够的有效样本量, 以保证推算能够满足精度的要求, 样本量也应适当放大。

6. 资源限制。调查项目的经费、时间要求及调查人员都是有限的, 因此样本量的确定也受这些因素的影响。

以上应考虑的问题有些属于非抽样误差研究的范围。

二、总体参数为 Y 或 Y 的情形

在简单随机抽样简单估计的情形下, 根据 y 的方差式

$$V(y) = \frac{1}{n} f S^2$$

代入

$$d = tS(\bar{y}) = t\sqrt{V(\bar{y})} = t\sqrt{\frac{N-n}{Nn}}S^2$$

或

$$r = tCv(\bar{y}) = t\sqrt{\frac{V(\bar{y})}{\bar{Y}^2}} = \frac{t}{\bar{Y}}\sqrt{\frac{N-n}{Nn}}S^2$$

得到

$$n = \frac{Nt^2S^2}{Nd^2 + t^2S^2} \quad (2.17)$$

$$\text{或 } n = \frac{Nt^2S^2}{Nr^2\bar{Y}^2 + t^2S^2} \quad (2.18)$$

在实际工作中,通常先计算

$$n_0 = \frac{t^2S^2}{d^2} \text{ 或 } n_0 = \frac{t^2S^2}{r^2\bar{Y}^2} = \left(\frac{t}{r}\right)^2 C^2 \quad (2.19)$$

式中, $C = \frac{S}{\bar{Y}}$ 为总体变异系数。如果 $\frac{n_0}{N} < 0.05$, 则就取 n_0 , 否则对 n_0 进行修正:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (2.20)$$

【例2.6】 在例2.3中,如果要求以95%的把握保证相对误差不超过10%,样本量应该至少是多少?

解:由该问题给出的条件: $N = 100, r = 10\% = 0.1$

置信度95%,对应的 $t = 1.96$

且已有 $\bar{y} = 5, s^2 = 19.1111$

计算样本量 n_0 :

$$n_0 = \left(\frac{t}{r}\right)^2 \frac{s^2}{\bar{y}^2} = \frac{1.96^2 \times 19.1111}{0.1^2 \times 5^2} \approx 294$$

计算修正样本量 n :

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{294}{1 + \frac{294}{100}} \approx 75$$

因此,应抽取一个大小至少为75的简单随机样本,才能满足95%置信度条件下相对误差不超过10%。

三、总体参数为 P 的情形

如果估计的是总体中某个特征的单元占总体的比例 P , 所用的估计量是样本

比例 p 时,则由

$$V(p) = \frac{PQ}{n} \frac{N-n}{N-1}$$

有

$$d = t \sqrt{V(p)} = t \sqrt{\frac{N-n}{N-1} \frac{PQ}{n}}$$

或

$$r = t \frac{\sqrt{V(p)}}{P} = \frac{t}{P} \sqrt{\frac{N-n}{N-1} \frac{PQ}{n}}$$

因此

$$n = \frac{t^2 \frac{PQ}{d^2}}{1 + \frac{1}{N} \left(t^2 \frac{PQ}{d^2} - 1 \right)} \quad (2.21)$$

或

$$n = \frac{t^2 \frac{Q}{r^2 P}}{1 + \frac{1}{N} \left(t^2 \frac{Q}{r^2 P} - 1 \right)} \quad (2.22)$$

在实际工作中,通常先计算

$$n_0 = \frac{t^2 PQ}{d^2} \text{ 或 } n_0 = \frac{t^2 Q}{r^2 P} \quad (2.23)$$

如果 $\frac{n_0}{N} < 0.05$, 就取 n_0 , 否则对 n_0 进行修正:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \quad (2.24)$$

在实际工作中,如果 P 在 0.5 附近,可根据 PQ 在 $P = Q = 0.5$ 时达到极大值来对样本量进行计算,这时将 t, d 以及 $PQ = 0.25$ 代入公式即可计算样本量。如果 $\frac{n_0}{N}$ 不能忽略,就对样本量进行必要的修正。例如,置信度为 95% 时(对应的 $t = 1.96$),最大允许绝对误差 $d = 1\%$,则必要的样本量为:

$$n_0 = \frac{t^2 PQ}{d^2} = \frac{1.96^2 \times 0.5 \times 0.5}{0.01^2} = 9604$$

若 $P < 0.1$ (或 $P > 0.9$),由于这时 PQ 与 0.25 相差太大,用 $PQ = 0.25$ 太过保守,这样计算的样本量太大。以 $d = 1\%, P = 0.1$ 为例,这时

$$n_0 = \frac{t^2 PQ}{d^2} = \frac{1.96^2 \times 0.1 \times 0.9}{0.01^2} \approx 3458$$

比9604要小很多。

【例2.7】某销售公司希望了解全部3000家客户对该公司的综合满意程度,决定用电话来调查一个简单随机样本。这时,销售公司希望以95%的把握保证客户满意的总体比例 P 在样本比例 $p \pm 10\%$ 的范围内,但对总体比例 P 无法给出一个大致范围。这时,应该调查多少个客户,才能保证对总体比例估计的要求?

解:由该问题给出的条件: $N = 3000, d = 10\% = 0.1$

置信度95%,对应的 $t = 1.96$

由于无法得到 P 的初始估计值,因此取使 PQ 达极大值的 $P = 0.5$,得到最保守的 n_0 :

$$n_0 = \frac{1.96^2 \times 0.5 \times 0.5}{0.1^2} \approx 96$$

计算修正样本量 n :

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} = \frac{96}{1 + \frac{96 - 1}{3000}} \approx 93$$

注意,在本问题中,由于 $\frac{n_0}{N} = \frac{96}{3000} = 3.2\% < 5\%$,所以修正样本量 n 与未修正样本量 n_0 差别不大,这也是抽样比小于5%时可以被忽略的原因。

四、总体参数的预先估计

由上述对样本量的计算可以看出,计算过程中, t, r 或 d 都可事先规定,但总体均值 \bar{Y} 以及总体方差 S^2 未知,这时需要对总体均值以及总体方差进行预估计。

在实际工作中,可以依照过去对同类问题调查的经验来估计。例如,对同类问题获得过一个样本量为 n_0 的简单随机样本,并且已知在一定置信度下(比如95%),该调查对总体均值(或总量)估计的相对误差为 r_0 ,则在同样的置信度下,如果希望本次调查的相对误差达到 r ,则在抽样比可忽略的情况下,可以近似地计算本次调查所需的样本量:

$$n = \frac{r_0^2}{r^2} n_0 \quad (2.23)$$

由这个公式看出,如果调查时希望相对误差减小到原来的一半,则所需的样本量为原来的4倍。例如 $r_0 = 20\%$,希望 $r = 10\%$,则 $n = 4n_0$ 。

有时,可通过预调查对总体均值及总体方差进行估计。一般来说,对于大型调

查,通常要进行预调查,预调查的目的主要是检查调查组织工作中可能出现哪些问题、问卷设计是否合理等,并加以解决。

有时,如果时间允许,且总体范围和目标量的数量特征不会随时间的变化有大的变化,调查可以分为两步。首先确定一个可以承受的样本量 n_0 ,调查后对估计精度进行计算,如果精度达到要求,则不再进行下一步;否则,计算为达到精度要求所需的样本量 n ,再调查 $n_1 = n - n_0$ 的补充样本。

有时,没有同类调查的经验,又不允许预调查,则只能通过定性分析,这时,最好是对总体变异系数 C 进行分析并估计,因为变异系数通常变化不大,根据对变异系数的估计,利用(2.19)对样本量进行计算。

相比较而言,如果估计的是总体比例 P ,只要根据分析确认 P 不是很稀有事件的比例,也即只要 P 在 $0.2 \sim 0.8$ 之间,问题就变得简单,因为这时可以取使 PQ 达到最大的 P 值(即 $P = 0.5$)来对样本量进行保守的估计。

§ 2.4 其他问题

一、逆抽样

如果估计的是稀有事件的比例,这时总体比例 P 很小,用相对误差 r 比绝对误差 d 更好些。试想,取 $d = 1\%$ 看上去很小,如果某个稀有事件的比例为 1% ,则实际上估计的精度很差,取 r 就能避免这种尴尬的情形。

对于稀有事件,所需的样本量会很大,我们来看看 $P_1 = 1\%$, $P_2 = 5\%$ 和 $P_3 = 10\%$ 时,在置信度 95% 的条件下,要达到 $r = 10\%$ 所分别需要的样本量(假定抽样比可忽略)。

$$P_1 = 1\% \text{ 时, } n_1 = \frac{t^2 Q_1}{r^2 P_1} = \frac{1.96^2 \times 0.99}{0.1^2 \times 0.01} = 38\,032$$

$$P_2 = 5\% \text{ 时, } n_2 = \frac{t^2 Q_2}{r^2 P_2} = \frac{1.96^2 \times 0.95}{0.1^2 \times 0.05} = 7\,299$$

$$P_3 = 10\% \text{ 时, } n_3 = \frac{t^2 Q_3}{r^2 P_3} = \frac{1.96^2 \times 0.9}{0.1^2 \times 0.1} = 3\,458$$

如果 P 为万分之一或十万分之一,所需的样本量更大。

对于稀有事件的比例估计问题,如果问题非常重要,的确需要按计算的样本量进行调查,问题在于,现在只知道要调查的是一个稀有事件,但无法给出它确切的范围,到底是万分之一还是十万分之一。从上面的例子可以看到,对总体比例事先

不同的假定,所导致的样本量差异非常大。遇到这种问题,可以采用霍丹(Haldane)提出的逆抽样方法,即事先确定一个整数 m ($m > 1$),进行逐个抽样,直至抽到 m 个所考虑特征的单元为止。设 n 是实际的样本量,则 P 的一个无偏估计为:

$$p = \frac{m}{n} - \frac{1}{m} \quad (2.25)$$

当 N 比较大, $m \geq 10$ 时,

$$V(p') \approx \frac{mP^2Q}{(m-1)^2} \quad (2.26)$$

从而估计量 p' 的变异系数为:

$$Cv(p') = \frac{S(p')}{P} \approx \frac{\sqrt{m}Q}{m-1} < \frac{\sqrt{m}}{m-1} \quad (2.27)$$

因为 Q 很接近于 1,因此 $\frac{\sqrt{m}}{m-1}$ 很接近于 $Cv(p')$ 的上界,由

$$r = tCv(p')$$

规定了 $Cv(p')$ 或 t, r 后,就可以确定 m 。如规定 $Cv(p') = 20\%$, 则 $m = 27$ 。

可以证明,这时所需样本量 n 的均值为:

$$E(n) = \frac{m}{P} \quad (2.28)$$

因此,对于非常稀有事件,实际上的样本量 n 是很大的。例如, P 为万分之一, $m = 27$, 则平均来看, $n = 270\,000$ 。

二、设计效应 (design effect)

为比较不同抽样方法的效率,我们可以通过抽样方法的设计效应(简称 deff)来进行比较。设计效应是由基什(L. Kish)提出的,其定义为:

$$deff = \frac{V(y)}{V_{rs}(y)} \quad (2.29)$$

式中, $V_{rs}(y)$ 为不放回简单随机抽样简单估计量的方差; $V(y)$ 为某个抽样设计在同样样本量条件下估计量的方差。

由设计效应的定义,它就是将某个抽样设计的估计量的方差与同样样本量条件下的不放回简单随机抽样简单估计量的方差进行比较。如果 $deff < 1$, 则所考虑的抽样设计比简单随机抽样的效率高;反之,如果 $deff > 1$, 则所考虑的抽样设计比简单随机抽样的效率低。

例如,放回简单随机抽样的 $deff$ 为:

$$deff = \frac{\frac{(N-1)S^2}{(Nn)}}{\frac{(N-n)S^2}{(Nn)}} = \frac{N}{N-1} \frac{1}{n}$$

显然,这时的 $deff > 1$,即放回简单随机抽样的效率比不放回简单随机抽样的效率低。

$deff$ 对复杂抽样时确定样本量有很大作用,在一定精度条件下,简单随机抽样所需的样本量 n' 比较容易得到,如果可以估计复杂抽样的 $deff$,那么复杂抽样所需的样本量为:

$$n = n' \times deff \quad (2.30)$$

小 结

本章介绍了简单随机抽样的理论及若干相关的问题。简单随机抽样的理论比较成熟,它是其他抽样方法的基础。可以说,其他抽样方法是在其基础上发展起来的。

在大多数情况下,简单随机抽样的效率比较高。但它的缺点是需要要在抽样之前编制一份完整的抽样框,并给抽样框中的每个单元赋予一个编号,使得简单随机抽样能够实施,这在实际工作中往往难以实现,因为当抽样的总体比较大时,编制一个完整的抽样框比较困难或者根本就不可能,这时就需要使用其他抽样方法。简单随机抽样的另一个缺点是样本在总体中比较分散,这往往使得调查难以实施,因为寻找样本单元可能比较困难或花费较多的时间,从而使得调查的费用大大提高,抽样调查费用节省的优点反而得不到体现。在实际工作中,如果出现这类问题,就必须采取其他抽样方法。

本章附录 简单随机抽样简单估计量性质的证明

放回简单随机抽样所得到的样本是独立同分布的样本,对其性质的证明在统计学教科书中可以找到。这里我们主要讨论不放回简单随机抽样估计量性质的证明。

由于总体总量、总体均值之间只差一个常数,即

$$Y = NY$$

对于总体中具有某种特征的单元的比例 P , 如果定义

$$Y_i = \begin{cases} 1, & \text{第 } i \text{ 个单元具有所考虑的特征} \\ 0, & \text{其他} \end{cases}, i = 1, 2, \dots, N$$

则有

$$P = \frac{A}{N} = \frac{1}{N} \sum_{i=1}^N Y_i = Y$$

对总体比例的估计类似于对总体均值的估计。

因此, 下面只证明性质 1、性质 2、性质 3, 分别对应的是 (2.2)、(2.5)、(2.6) 式。

1. 证明性质 1: 对于简单随机抽样, y 是 Y 的无偏估计。

证明: 这个性质的证明方法有多种, 下面介绍其中的两种。

方法一: 对于固定的有限总体, 估计量的期望是对所有可能样本求平均得到的。对于一个大小为 N 的总体, 所有样本量为 n 的简单随机样本有 C_N^n 个, 因此

$$E(y) = \frac{\sum y}{C_N^n} = \frac{\sum (y_1 + y_2 + \dots + y_n)}{nC_N^n}$$

式中的求和号是对所有 C_N^n 个样本的求和。为了求出分子上的和式, 我们要算出总体中每个特定的单元 y_i 在不同的样本中出现的次数。由于当样本中含有特定的单元 y_i 时, 样本中其他 $n-1$ 个位置的单元要从总体 $N-1$ 个单元中抽取, 因此, 含有 y_i 的样本共有 C_{N-1}^{n-1} 个, 于是分子为:

$$\sum y = \frac{1}{n} \sum (y_1 + y_2 + \dots + y_n) = \frac{1}{n} C_{N-1}^{n-1} \sum_{i=1}^N Y_i$$

注意到

$$C_N^n = \frac{N!}{n!(N-n)!} = \frac{N}{n} \frac{(N-1)!}{(n-1)!(N-n)!} = \frac{N}{n} C_{N-1}^{n-1}$$

所以

$$E(y) = \frac{\sum y}{C_N^n} = \frac{C_{N-1}^{n-1} \sum_{i=1}^N Y_i}{nC_N^n} = \frac{1}{N} \sum_{i=1}^N Y_i = Y$$

方法二: (对称论证法) 由于每个单元出现在总体所有可能样本中的次数相同, 因此

$$E(y_1 + y_2 + \dots + y_n)$$

一定是

$$Y_1 + Y_2 + \dots + Y_N$$

的倍数,且这个倍数就是 $\frac{n}{N}$,因为前者有 n 项,而后者是 N 项,因此

$$E(y) = \frac{1}{n}E\left(\sum_{i=1}^n y_i\right) = \frac{1}{n} \cdot \frac{n}{N} \sum_{i=1}^N Y_i = \bar{Y}$$

2. 证明性质 2: 对于简单随机抽样, y 的方差为 $V(y) = \frac{1}{n} \left(1 - \frac{n}{N}\right) S^2$.

证明: 由定义

$$\begin{aligned} V(y) &= E(y - \bar{Y})^2 = E\left(\frac{1}{n} \sum_{i=1}^n y_i - \bar{Y}\right)^2 \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^n (y_i - \bar{Y})\right]^2 \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^n (y_i - \bar{Y})^2\right] + \frac{1}{n^2} E\left[\sum_{i \neq j} (y_i - \bar{Y})(y_j - \bar{Y})\right] \end{aligned}$$

根据对称论证法,有

$$E\left[\sum_{i=1}^n (Y_i - \bar{Y})^2\right] = \frac{n}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2$$

由于 $E\left[\sum_{i \neq j} (y_i - \bar{Y})(y_j - \bar{Y})\right]$ 中的求和是对 $\frac{n(n-1)}{2}$ 项的, $\sum_{i \neq j} (Y_i - \bar{Y})(Y_j - \bar{Y})$ 中求和是对 $\frac{N(N-1)}{2}$ 项的,因此,根据对称论证法,有

$$E\left[\sum_{i \neq j} (y_i - \bar{Y})(y_j - \bar{Y})\right] = \frac{n(n-1)}{N(N-1)} \sum_{i \neq j} (Y_i - \bar{Y})(Y_j - \bar{Y})$$

因此

$$\begin{aligned} V(y) &= \frac{1}{n^2} E\left[\sum_{i=1}^n (y_i - \bar{Y})^2\right] + \frac{1}{n^2} E\left[\sum_{i \neq j} (y_i - \bar{Y})(y_j - \bar{Y})\right] \\ &= \frac{1}{n^2} \cdot \frac{n}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 + \frac{1}{n^2} \cdot \frac{n(n-1)}{N(N-1)} \sum_{i \neq j} (Y_i - \bar{Y})(Y_j - \bar{Y}) \\ &= \frac{1}{nN} \left[\sum_{i=1}^N (Y_i - \bar{Y})^2 + \frac{n-1}{N-1} \sum_{i \neq j} (Y_i - \bar{Y})(Y_j - \bar{Y}) \right] \\ &= \frac{1}{nN} \left\{ \left(1 - \frac{n-1}{N-1}\right) \sum_{i=1}^N (Y_i - \bar{Y})^2 + \frac{n-1}{N-1} \left[\sum_{i=1}^N (Y_i - \bar{Y})\right]^2 \right\} \\ &= \frac{1}{nN} \cdot \frac{N-n}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{N-n}{nN} S^2 \end{aligned}$$

$$\frac{1-f}{n} S^2$$

3. 证明性质 3: $V(y)$ 的样本无偏估计为 $v(y) = \frac{1-f}{n} s^2$.

证明: 将 s^2 改写为:

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2 \right] \end{aligned}$$

由对称论证法

$$E \left[\sum_{i=1}^n (y_i - \bar{Y})^2 \right] = \frac{n}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{n(N-1)}{N} S^2$$

根据性质 2, 有

$$E(\bar{y} - \bar{Y})^2 = \frac{1-f}{n} S^2 = \frac{N-n}{nN} S^2$$

因此

$$\begin{aligned} E(s^2) &= \frac{1}{n-1} \{ E \left[\sum_{i=1}^n (y_i - \bar{Y})^2 \right] - nE(\bar{y} - \bar{Y})^2 \} \\ &= \frac{1}{n-1} \left[\frac{n(N-1)}{N} S^2 - n \frac{N-n}{nN} S^2 \right] \\ &= \frac{S^2}{N(n-1)} [n(N-1) - (N-n)] \\ &= S^2 \end{aligned}$$

由于简单随机样本的方差 s^2 是总体方差 S^2 的无偏估计, 因此 $v(\bar{y})$ 是 $V(y)$ 的无偏估计。

习 题

1. 判断下面要估计的总体目标量分别属于什么类型:

- (1) 测试一名大学生的英语词汇量;
- (2) 调查城市居民家庭平均用电量;
- (3) 估计空气中氮氧化合物的含量;
- (4) 估计湖中鱼的数量;
- (5) 测试日光灯的寿命;

(6) 估计居民家庭用于做饭菜及饮用的用水量占家庭总用水量的比重;

(7) 估计婴儿出生性别比;

(8) 检测食盐中的碘含量

2. 试判断下面数字的产生是否是等概率的

(1) 总体编号为 1 ~ 35, 在 00 ~ 99 中产生随机数 r , 若 $r = 00$ 或 $r > 35$, 则抛弃重抽.

(2) 总体编号为 1 ~ 35, 在 00 ~ 99 中产生随机数 r , 若 $r \geq 50$, 则取 $r' = r - 50$, 否则 $r' = r$. 如果 $r' = 00$ 或 $r' > 35$, 则抛弃重抽.

(3) 总体编号为 1 ~ 35, 在 00 ~ 99 中产生随机数 r , 以 r 除以 35, 余数作为被抽中的数, 如果余数为 0, 则被抽中的数为 35.

(4) 总体编号为 1 580 ~ 2 107, 在 000 ~ 527 中产生一个随机数, 以 $1\,580 + r$ 为被抽中的数.

3. 某项长期调查项目在全面展开之前进行了试点, 调查了一个样本量为 800 的简单随机样本, 方案设计人员以这个样本为总体, 计算出达到精度要求时需要的样本量为 80, 从而相应的抽样比为 10%, 据此, 方案设计人员要求在以后的调查中, 抽样比为 10%, 即必须调查总体单位中的 10%. 你认为设计人员的做法有何不妥?

4. 设总体为 0, 1, 3, 5, 6, 计算总体均值 Y 、总体方差 σ^2 和 S^2 ; 给出全部 $n = 2$ 的样本, 并验证 $E(\bar{y}) = Y$ 及 $E(s^2) = S^2$.

5. 为调查学生购书支出, 某高校在令校 6 000 名大学生中按简单随机抽样抽出 78 名学生, 调查了他们最近一个学期用于购书支出后, 得到 $\bar{y} = 102.30$ (元), $s^2 = 13\,712$, 试估计该校大学生最近一个学期用于购书的总支出, 并给出估计的标准差. 若要求在置信度 95% 下, 估计的相对误差不超过 10%, 则应该抽出多少学生进行调查?

6. 从 5 000 个电子元器件的一批产品中无放回地随机抽取 100 个并进行了检验, 其中合格品为 93 个, 试估计这批产品的合格率, 并给出估计的标准差. 如果在 95% 置信度下, 要使估计的绝对误差不超过 1%, 则需要多少样本量?

7. 为测试学习某种手工操作所需的时间, 在人群中随机抽选了 10 名志愿者, 记录下 10 名志愿者掌握这项操作所用的时间(单位: 分): 16, 21, 17, 15, 26, 20, 24, 23, 24, 21, 试估计学习该手工操作所用的平均时间, 并给出估计 95% 的置信区间.

8. 某地区拥有 10 万户居民, 某保险公司欲对该地区居民购买保险的情况进行调查, 在全体居民户中按简单随机抽样抽出 50 户居民户, 通过调查得知其中有 3 户购买了保险. 试估计该地区居民户投保的比例, 并给出估计的标准差. 如果希望

在 95% 置信度下(对应的 $t = 2$),估计的绝对误差不超过 1%,则所需的样本量为多少?

9. 对某个问题获得了一个样本量为 n_0 的简单随机样本,在一定置信度下(对应 t),该样本对总体均值估计的相对误差为 r_0 ,则在同样的置信度下,如果希望相对误差达到 r ,则在抽样比可忽略的情况下,证明所需的样本量为:

$$n = \frac{r_0^2}{r^2} n_0$$



第 3 章

分层随机抽样

学习了简单随机抽样之后,我们知道了影响估计精度的因素除了样本量、总体大小(通常不是主要因素)以外,还有总体的方差。也就是说在其他因素不变的情况下,总体方差越大,估计的精度越差;反之,估计的精度就越高。对于一个总体,其方差是客观存在且无法改变的,但如果对总体单元进行分类,即分成若干子总体,在子总体内单元之间比较相似,使每一个子总体的方差变小,这样只需在子总体中抽取少量样本单元,就能很好地代表子总体的特征,从而提高对整个总体估计的精度。这就是人们常用的分层抽样技术。

本章共分五节,第一节将介绍分层随机抽样的定义、使用场合以及符号;第二节介绍估计量及其性质;第三节介绍样本量的分配原则;第四节介绍样本量的确定;第五节介绍分层抽样的若干问题。

§ 3.1 引 言

一、定义与作用

(一) 定义

在抽样之前,先将总体 N 个单元划分成 L 个互不重复的子总体,每个子总体

称为层,它们的大小分别为 N_1, N_2, \dots, N_L , 这 L 个层合起来就是整个总体 ($N = \sum_{h=1}^L N_h$) 然后,在每个层中分别独立地进行抽样,这种抽样就是分层抽样,所得到的样本称为分层样本。如果每层都是简单随机抽样,则称为分层随机抽样,所得到的样本称为分层随机样本。

上述定义也表明,总体中的每一个单元一定属于并且只属于某一个层,而不可能同时属于两个层或不属于任何一个层。

(二) 作用

分层抽样在实际工作中应用得非常广泛,主要是因为它具有其他抽样方法所没有的特点。

1. 分层抽样的抽样效率较高,也就是说分层抽样的估计精度较高。这是因为分层抽样估计量的方差只和层内方差有关,和层间方差无关。因此,人们可以通过对总体分层,尽可能地降低层内差异,使层间差异大,从而提高估计的精度。另外,直观上也可以想像得出,简单随机抽样可能出现极端的情况,样本偏向某一部分,而分层抽样每层都要抽取一定的样本单元,因此样本在总体中分布比较均匀。

2. 分层抽样不仅能对总体指标进行推算,而且能对各层指标进行推算。有时调查的目的不仅要推算总体指标,可能还要推算各层的指标。例如,某市对全市企业进行抽样调查,要求最终能给出各行业的指标,因此按行业分层后,所得的样本不仅能推算全市的指标,也能对各行业进行推算。

3. 层内抽样方法可以不同,而且便于抽样工作的组织。例如,某项全国范围的大型抽样调查,要编制全国范围的抽样框往往是一件非常困难的事,但如果抽样按行政区划或行业分层后,可以调动各级主管部门的积极性,分头编制抽样框并实施抽样的组织和调查工作。为了组织调查的方便,各层可以根据层内的特点,分别采用不同的抽样方法。

二、使用场合

根据分层抽样的特点,分层除了可以提供子总体指标和便于调查的组织实施,通常,使用分层抽样的主要目的是为了提高估计的精度。为充分利用分层抽样的特点,在一项抽样调查项目中,往往反复使用分层抽样方法。

在对层进行具体划分时,通常考虑如下原则:

1. 层内单元具有相同性质,通常按调查对象的不同类型进行划分。这时,分层抽样能够对每一类的目标量进行估计。

2. 尽可能使层内单元的标志值相近,层间单元的差异尽可能大,从而达到提

高抽样估计精度的目的。

3. 既按类型又按层内单元标志值相近的原则进行多重分层,同时达到实现估计类值以及提高估计精度的目的。

4. 抽样组织实施的方便,通常按行政管理机构设置进行分层。

通常用于分层的指标有行政区划、地理位置、海拔高度、行业、经济发达程度、企业规模、家庭收入水平、性别等。

例如,对全国范围汽车运输的抽样调查,调查目的不仅要推算全国货运汽车完成的运量,还要推算不同经济成分(国有、集体、个体)汽车完成的运量。为组织的方便,首先将货运汽车总体按省分层,由各省运输管理部门负责省内的调查工作;各省再将省内拥有的汽车按经济成分分层;为提高抽样效率,再按吨位对汽车分层。

又如,某高校对学生在宿舍使用电脑的情况进行调查,根据经验,本科生和研究生拥有电脑的状况差异较大,因此,在抽样前对学生按本科生和研究生进行分层是有必要的。

三、符号说明

我们用下标 h 表示层号($h = 1, 2, \dots, L$)。关于第 h 层的记号如下:

单元总数: N_h

样本单元数: n_h

第 i 个单元标志值(观察值): y_{hi}

层权: $W_h = \frac{N_h}{N}$

抽样比: $f_h = \frac{n_h}{N_h}$

总体均值: $Y_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$

样本均值: $y_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$

总体方差: $S_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (Y_{hi} - Y_h)^2$

样本方差: $s_h^2 = \frac{1}{n_h} \sum_{i=1}^{n_h} (y_{hi} - y_h)^2$

§ 3.2 估计量

一、总体均值的估计

(一) 简单估计量的定义

对于分层样本,对总体均值 Y 的估计是通过对各层的 Y_h 的估计,按层权 W_h 加权平均得到的 公式为:

$$\hat{Y}_s = \sum_{h=1}^L W_h \hat{Y}_h = \frac{1}{N} \sum_{h=1}^L N_h \hat{Y}_h \quad (3.1)$$

如果得到的是分层随机样本,则总体均值 Y 的简单估计为:

$$y_s = \sum_{h=1}^L W_h y_h = \frac{1}{N} \sum_{h=1}^L N_h y_h \quad (3.2)$$

(二) 估计量的性质

性质 1 对于一般的分层抽样,如果 \hat{Y}_h 是 Y_h 的无偏估计($h = 1, 2, \dots, L$),则 \hat{Y}_s 是 Y 的无偏估计, \hat{Y}_s 的方差为:

$$V(\hat{Y}_s) = \sum_{h=1}^L W_h^2 V(\hat{Y}_h) \quad (3.3)$$

值得注意的是,只要对各层估计是无偏的,则对总体的估计也是无偏的。因此,各层可以采用不同的抽样方法,只要相应的估计量是无偏的,则对总体的推算也是无偏的。

性质 2 对于分层随机抽样, y_s 是 Y 的无偏估计, y_s 的方差为:

$$V(y_s) = \sum_{h=1}^L W_h^2 V(y_h) = \sum_{h=1}^L W_h^2 \frac{1-f_h}{n_h} S_h^2 \quad (3.4)$$

性质 3 对于分层随机抽样, $V(y_s)$ 的一个无偏估计为:

$$v(y_s) = \sum_{h=1}^L W_h^2 v(y_h) = \sum_{h=1}^L W_h^2 \frac{1-f_h}{n_h} s_h^2 \quad (3.5)$$

二、总体总量的估计

(一) 简单估计量的定义

总体总量 Y 的估计为:

$$\hat{Y} = N\hat{Y}_q = \sum_{h=1}^L \hat{Y}_h \quad (3.6)$$

如果得到的是分层随机样本,则总体总量 Y 的简单估计为:

$$\hat{Y} = N\hat{y}_q \quad (3.7)$$

(二) 估计量的性质

由于 \hat{Y} 与 \hat{Y}_q 只差一个常数,因此, \hat{Y} 与 \hat{Y}_q 具有同样的性质。

性质4 对于一般的分层抽样,如果 \hat{Y}_q 是 Y 的无偏估计,则 \hat{Y} 是 Y 的无偏估计。 \hat{Y} 的方差为:

$$\begin{aligned} V(\hat{Y}) &= N^2 V(\hat{Y}_q) = \sum_{h=1}^L V(\hat{Y}_h) \\ &= N^2 \sum_{h=1}^L W_h^2 V(\hat{Y}_h) = \sum_{h=1}^L N_h^2 V(\hat{Y}_h) \end{aligned} \quad (3.8)$$

性质5 对于分层随机抽样, \hat{Y} 的方差为:

$$V(\hat{Y}) = \sum_{h=1}^L N_h^2 V(\hat{Y}_h) = \sum_{h=1}^L N_h^2 \frac{1}{n_h} f_h S_h^2 \quad (3.9)$$

性质6 对于分层随机抽样, $V(\hat{Y})$ 的一个无偏估计为:

$$v(\hat{Y}) = \sum_{h=1}^L N_h^2 v(y_h) = \sum_{h=1}^L N_h^2 \frac{1}{n_h} f_h s_h^2 \quad (3.10)$$

【例3.1】 调查某地区的居民奶制品年消费支出,以居民户为抽样单元,根据经济及收入水平将居民户划分为4层,每层按简单随机抽样抽取10户,调查获得如下数据(单位:元),如表3.1,估计该地区居民奶制品年消费总支出及估计的标准差。

表 3.1 样本户奶制品年消费支出

层	居民户总数	样本户奶制品年消费支出									
		1	2	3	4	5	6	7	8	9	10
1	200	10	40	0	110	15	10	40	80	90	0
2	400	50	130	60	80	100	55	160	85	160	170
3	750	180	260	110	0	140	60	200	180	300	220
4	1500	50	35	15	0	20	30	25	10	30	25

解:由上表, $N = 2850$, $n_h = 10 (h = 1, 2, 3, 4)$

各层的层权及抽样比为:

$$\begin{aligned}
W_1 &= \frac{N_1}{N} = \frac{200}{2\,850} \approx 0.070\,18 & f_1 &= \frac{n_1}{N_1} = \frac{10}{200} \approx 0.05 \\
W_2 &= \frac{N_2}{N} = \frac{400}{2\,850} \approx 0.140\,35 & f_2 &= \frac{n_2}{N_2} = \frac{10}{400} \approx 0.025 \\
W_3 &= \frac{N_3}{N} = \frac{750}{2\,850} \approx 0.263\,16 & f_3 &= \frac{n_3}{N_3} = \frac{10}{750} \approx 0.013\,3 \\
W_4 &= \frac{N_4}{N} = \frac{1\,500}{2\,850} \approx 0.526\,32 & f_4 &= \frac{n_4}{N_4} = \frac{10}{1\,500} \approx 0.006\,7
\end{aligned}$$

各层样本均值及样本方差为:

$$y_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} = 39.5$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_{1i} - y_1)^2 \approx 1\,624.722$$

同理可得 $y_2 = 105, y_3 = 165, y_4 = 24$

$$s_2^2 \approx 2\,166.667, s_3^2 \approx 8\,205.556, s_4^2 \approx 193.333$$

因此,估计奶制品年消费总支出为:

$$\begin{aligned}
\hat{Y} &= \sum_{h=1}^4 N_h y_h \\
&= 200 \times 39.5 + 400 \times 105 + 750 \times 165 + 1\,500 \times 24 \\
&= 209\,650 (\text{元})
\end{aligned}$$

估计量方差及标准差的样本估计为:

$$v(\hat{Y}) = N^2 \sum_{h=1}^4 W_h^2 v(y_h) = \sum_{h=1}^4 N_h^2 \frac{1-f_h}{n_h} s_h^2 \approx 5.39 \times 10^8$$

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} \approx 23\,208 (\text{元})$$

因此,我们可以以 95% 的把握说该地区居民奶制品年消费总支出在

$$\hat{Y} \pm t_s(\hat{Y}) = 209\,650 \pm 1.96 \times 23\,208 (\text{元})$$

之间。换句话说,在 164\,162 元 ~ 255\,138 元之间

三、总体比例的估计

(一) 简单估计量的定义

总体比例 P 的估计为:

$$p_d = \sum_{h=1}^H W_h p_h \quad (3.11)$$

(二) 估计量的性质

如果定义

$$Y_i = \begin{cases} 1, & \text{第 } i \text{ 个单元具有所考虑的特征} \\ 0, & \text{其他} \end{cases}, i = 1, 2, \dots, N$$

则对总体比例的估计类似对总体均值的估计, 这时 p_{st} 与 \bar{Y}_{st} 具有同样的性质。

性质 7 对于一般的分层抽样, 如果 p_h 是 P_h 的无偏估计 ($h = 1, 2, \dots, L$), 则 p_{st} 是 P 的无偏估计, p_{st} 的方差为:

$$V(p_{st}) = \sum_{h=1}^L W_h^2 V(p_h) \quad (3.12)$$

性质 8 对于分层随机抽样, p_{st} 是 P 的无偏估计, 注意到

$$V(p_h) = \frac{N_h - n_h}{N_h - 1} \frac{P_h Q_h}{n_h} \text{ 及 } N_h - 1 \approx N_h$$

因而 p_{st} 的方差为:

$$\begin{aligned} V(p_{st}) &= \sum_{h=1}^L W_h^2 V(p_h) = \frac{1}{N^2} \sum_{h=1}^L \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h} \\ &\approx \sum_{h=1}^L \frac{1}{N^2} \frac{N_h^2 (N_h - n_h)}{N_h} \frac{P_h Q_h}{n_h} \\ &= \sum_{h=1}^L W_h^2 (1 - f_h) \frac{P_h Q_h}{n_h} \end{aligned} \quad (3.13)$$

性质 9 对于分层随机抽样, $V(p_{st})$ 的一个无偏估计为:

$$\begin{aligned} v(p_{st}) &= \sum_{h=1}^L W_h^2 v(p_h) = \frac{1}{N^2} \sum_{h=1}^L \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{p_h q_h}{n_h - 1} \\ &\quad \sum_{h=1}^L W_h^2 (1 - f_h) \frac{p_h q_h}{n_h - 1} \end{aligned} \quad (3.14)$$

【例 3.2】 在例 3.1 的调查中, 同时调查了居民户拥有家庭电脑的情况, 获得如下数据(单位: 台), 如表 3.2。估计该地区居民拥有家庭电脑的比例及估计的标准差。

表 3.2 样本户拥有家庭电脑情况

层	居民户总数	样本户拥有家庭电脑情况									
		1	2	3	4	5	6	7	8	9	10
1	200	0	0	0	1	0	0	0	1	0	0
2	400	0	1	0	0	0	0	0	0	1	0
3	750	1	1	0	0	0	0	1	0	1	0
4	1 500	1	0	0	0	0	0	0	0	0	0

解:由上表可得

$$p_1 = 0.2, p_2 = 0.2, p_3 = 0.4, p_4 = 0.1$$

根据前面对各层层权 W_h 及抽样比 f_h 的计算结果,可得各层估计量的方差:

$$v(p_1) = (1 - f_1) \frac{p_1 q_1}{n_1 - 1} \approx 0.0169$$

$$v(p_2) = (1 - f_2) \frac{p_2 q_2}{n_2 - 1} \approx 0.0173$$

$$v(p_3) = (1 - f_3) \frac{p_3 q_3}{n_3 - 1} \approx 0.0263$$

$$v(p_4) = (1 - f_4) \frac{p_4 q_4}{n_4 - 1} \approx 0.0099$$

因此,该地区居民拥有家庭电脑比例的估计为:

$$\begin{aligned} p_x &= \sum_{h=1}^4 W_h p_h = \frac{1}{N} \sum_{h=1}^4 N_h p_h \\ &= \frac{1}{2850} (200 \times 0.2 + 400 \times 0.2 + 750 \times 0.4 + 1500 \times 0.1) \\ &= 0.2 \end{aligned}$$

估计量的方差为:

$$\begin{aligned} v(p_x) &= \frac{1}{N^2} \sum_{h=1}^4 N_h^2 v(p_h) \\ &= \frac{1}{2850^2} (200^2 \times 0.0169 + 400^2 \times 0.0173 + 750^2 \times 0.0263 \\ &\quad + 1500^2 \times 0.0099) \\ &\approx 0.005 \end{aligned}$$

估计量的标准差为:

$$s(p_x) = \sqrt{v(p_x)} \approx 0.07$$

§ 3.3 样本量在各层的分配

对于分层抽样,当总的样本量一定时,还需研究各层应该分配多少样本量的问题,因为对总体推算时,估计量的方差不仅与各层的方差有关,还与各层所分配的样本量有关。实际工作中有不同的分配方法,可以按各层单元数占总体单元数的比例分配,也可以采用使估计量总方差达到最小等几种方法进行样本量的分配。

一、比例分配

这里的比例分配指的是按各层单元数占总体单元数的比例,也就是按各层的层权进行分配,这时

$$\frac{n_h}{n} = \frac{N_h}{N} = W_h \text{ 或 } f_h = \frac{n_h}{n} = \frac{N_h}{N} = f \quad (3.15)$$

对于分层随机抽样,这时总体均值 Y 的估计是:

$$\begin{aligned} y_{prop} &= \sum_{h=1}^L W_h y_h = \sum_{h=1}^L \frac{n_h}{n} y_h = \sum_{h=1}^L \frac{n_h}{n} \cdot \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \\ &= \frac{1}{n} \sum_{h=1}^L \sum_{i=1}^{n_h} y_{hi} = \frac{1}{n} \sum_{i=1}^N y_i = \bar{y} \end{aligned} \quad (3.16)$$

总体比例 P 的估计是:

$$p_{prop} = p = \frac{1}{n} \sum_{h=1}^L a_h \quad (3.17)$$

这是因为总体中的任一个单元,不管它在哪一个层,都以同样的概率入样,因此按比例分配的分层随机样本,估计量的形式特别简单。这种样本也称为自加权的样本。

y_{prop} 的方差为:

$$\begin{aligned} V(y_{prop}) &= \sum_{h=1}^L W_h^2 V(y_h) = \sum_{h=1}^L W_h^2 \frac{n_h}{n} \frac{1-f_h}{n_h} S_h^2 \\ &= \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2 \end{aligned} \quad (3.18)$$

p_{prop} 的方差为:

$$V(p_{prop}) = \frac{1-f}{Nn} \sum_{h=1}^L \frac{N_h^2 P_h Q_h}{N_h - 1} \approx \frac{1-f}{n} \sum_{h=1}^L W_h P_h Q_h \quad (3.19)$$

二、最优分配

(一) 最优分配

在分层随机抽样中,如何将样本量分配到各层,使得在总费用给定的条件下,估计量的方差达到最小,或在给定估计量方差的条件下,使总费用最小,能满足这个条件的样本量分配就是最优分配。

如果我们考虑简单线性费用函数,总费用

$$C = c_1 + \sum_{h=1}^L c_h n_h \quad (3.20)$$

则这时的最优分配是:

$$n_h = n \frac{W_h S_h}{\sum_{h=1}^L W_h S_h} = n \frac{N_h S_h}{\sum_{h=1}^L N_h S_h}, h = 1, 2, \dots, L \quad (3.21)$$

由此得出下面的行为准则,如果某一层单元数较多,内部差异较大,费用比较省,则对这一层的样本量要多分配一些。

(二) Neyman(内曼)分配

对于分层随机样本,作为特例,如果每层抽样的费用相同,即 $c_h = c$ 时,最优分配可简化为:

$$n_h = n \frac{W_h S_h}{\sum_{h=1}^L W_h S_h} = n \frac{N_h S_h}{\sum_{h=1}^L N_h S_h}, h = 1, 2, \dots, L \quad (3.22)$$

这种分配称为 Neyman 分配。这时, $V(y_q)$ 达到最小

$$V_{\min}(y_q) = \frac{1}{n} \left(\sum_{h=1}^L W_h S_h \right)^2 = \frac{1}{N} \sum_{h=1}^L W_h S_h^2 \quad (3.23)$$

【例 3.3】(续例 3.1) 如果样本量仍为 $n = 40$, 则按比例分配和 Neyman 分配时,各层的样本量应为多少?

解:按比例分配时,各层的样本量为:

$$n_1 = W_1 n \approx 0.07018 \times 40 = 2.81$$

$$n_2 = W_2 n \approx 0.14035 \times 40 = 5.61$$

$$n_3 = W_3 n \approx 0.26316 \times 40 = 10.53$$

$$n_4 = W_4 n \approx 0.52632 \times 40 = 21.05$$

即各层的样本量分别为 3, 6, 11, 20

对于 Neyman 分配,根据前面对 W_h 及 s_h 的计算结果,得到

$$W_1 s_1 \approx 0.07018 \times \sqrt{1624.722} = 2.8286$$

$$W_2 s_2 \approx 0.14035 \times \sqrt{2166.667} = 6.5330$$

$$W_3 s_3 \approx 0.26316 \times \sqrt{8205.556} = 23.8380$$

$$W_4 s_4 \approx 0.52632 \times \sqrt{193.333} = 7.3181$$

$$\sum_{h=1}^L W_h s_h = 2.8286 + 6.5330 + 23.8380 + 7.3181 = 40.51775$$

因此,按 Neyman 分配时,各层应分配的样本量为:

$$\begin{aligned} n_1 &= n \frac{W_1 s_1}{\sum_{h=1}^L W_h s_h} = 40 \times \frac{2.8286}{40.51775} \approx 2.79 \\ n_2 &\approx 6.45 \\ n_3 &\approx 23.53 \\ n_4 &\approx 7.23 \end{aligned}$$

即各层的样本量分别为 3, 7, 23, 7。

(三) 某些层要求大于 100% 抽样时的修正

按最优分配时,有时抽样比 $f = \frac{n}{N}$ 较大,某个层的 S_h 又比较大,则可能出现按最优分配计算的这个层的样本量 n_h 超过 N_h 的情况。实际工作中,如果第 k 层出现这种情况,最优分配是对这个层进行 100% 的抽样,即取 $n_k = N_k$,然后,将剩下的样本量 $n - n_k$ 按最优分配分到各层

§ 3.4 样本量的确定

一、一般公式

令 $n_h = n w_h$, 其中 w_h 已经选定,于是当方差 V 给定时,由式(3.4):

$$\begin{aligned} V &= \sum_{h=1}^L W_h^2 \frac{1-f_h}{n_h} S_h^2 = \sum_{h=1}^L \frac{W_h^2}{n_h} S_h^2 = \sum_{h=1}^L \frac{W_h^2}{N_h} S_h^2 \\ &= \frac{1}{n} \sum_{h=1}^L \frac{W_h^2}{w_h} S_h^2 = \frac{1}{N} \sum_{h=1}^L \frac{W_h^2}{w_h} S_h^2 \end{aligned}$$

得到确定样本量的一般公式为:

$$n = \frac{\sum_{h=1}^L \frac{W_h^2 S_h^2}{w_h}}{V + \frac{\sum_{h=1}^L W_h S_h^2}{N}} \quad (3.24)$$

如果估计精度是以误差限的形式给出,则 $V = \left(\frac{d}{t}\right)^2 = \left(\frac{r\bar{Y}}{t}\right)^2$, d 为绝对误差限, r 为相对误差限; t 为标准正态分布的双侧 α 分位数; \bar{Y} 为总体均值。这时,上式也可以表示为:

$$n = \frac{\sum W_h^2 S_h^2}{\left(\frac{d}{t}\right)^2 + \frac{\sum W_h S_h^2}{N}} = \frac{\sum W_h^2 S_h^2}{\left(\frac{rY}{t}\right)^2 + \frac{\sum W_h S_h^2}{N}} \quad (3.25)$$

当按比例分配时, $w_h = W_h$

$$n = \frac{\sum W_h S_h^2}{V + \frac{\sum W_h S_h^2}{N}} \quad (3.26)$$

实际工作中, n 的计算可以分为两步, 先计算

$$n_0 = \frac{\sum W_h S_h^2}{V}$$

然后进行修正:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

当按 Neyman 分配时, $w_h = \frac{W_h S_h}{\sum W_h S_h}$

$$n = \frac{(\sum W_h S_h)^2}{V + \frac{\sum W_h S_h^2}{N}} \quad (3.27)$$

【例 3.4】(续例 3.1) 如果要求在 95% 置信度下, 相对误差不超过 10%, 则按比例分配和 Neyman 分配时, 总样本量分别为多少?

解: 当按比例分配时:

由前面的计算结果, 可以得到各层的 $W_h s_h^2$

$$W_1 s_1^2 = \frac{N_1 s_1^2}{N} = \frac{200}{2\,850} \times 1\,624.722 \approx 114.016$$

$$W_2 s_2^2 = \frac{N_2 s_2^2}{N} = \frac{400}{2\,850} \times 2\,166.667 \approx 304.094$$

$$W_3 s_3^2 = \frac{N_3 s_3^2}{N} = \frac{750}{2\,850} \times 8\,205.556 \approx 2\,159.36$$

$$W_4 s_4^2 = \frac{N_4 s_4^2}{N} = \frac{1\,500}{2\,850} \times 193.333 \approx 101.754$$

$$\sum W_h s_h^2 = 2\,679.22$$

在 95% 置信度时, 对应的 $t = 1.96$, 又 $y_q = \frac{\hat{Y}}{N} = \frac{209\,650}{2\,850} \approx 73.561\,4$

因此得到

$$V = \left(\frac{\sigma_d}{t} \right)^2 = \left(\frac{0.1 \times 73.5614}{1.96} \right)^2 = 14.086$$

由此可以得到

$$n = \frac{\sum W_h s_h^2}{V} = \frac{2679.22}{14.086} \approx 190.2$$

对 n , 进行修正, 得到修正后的 n :

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{190.2}{1 + \frac{190.2}{2850}} \approx 179$$

当按 Neyman 分配时:

由前面的计算, 已有 V , 各层的 $W_h s_h$, $W_h s_h^2$ 及 $\sum W_h s_h$, $\sum W_h s_h^2$ 。

因此, 按 Neyman 分配时所需样本量 n 为:

$$n = \frac{(\sum W_h s_h)^2}{V + \frac{\sum W_h s_h^2}{N}} = \frac{40.51775^2}{14.086 + \frac{2679.22}{2850}} \approx 110$$

综合上述, 按比例分配时, 样本量至少应为 179; 按 Neyman 分配时, 样本量至少应为 110

二、最优分配需要考虑费用时

在最优分配时, 如果考虑费用为简单线性费用函数

$$C = c_0 + \sum_{h=1}^L c_h n_h$$

则由式(3.21):

$$w_h = \frac{\frac{W_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^L \frac{W_h S_h}{\sqrt{c_h}}}, h = 1, 2, \dots, L$$

当方差 V 给定时, 将其代入式(3.24) 得到样本量为:

$$n = \frac{(\sum W_h S_h \sqrt{c_h})^2}{V + \frac{\sum W_h S_h^2}{N}} \quad (3.28)$$

而当总费用 C 是给定时, 由本章附录 4, 有

$$n_h = K \frac{W_h S_h}{\sqrt{c_h}}$$

则

$$C - c_0 = \sum_{h=1}^L c_h n_h = K \sum_{h=1}^L \sqrt{c_h} W_h S_h$$

$$K = \frac{C - c_0}{\sum_{h=1}^L \sqrt{c_h} W_h S_h}$$

$$n_h = \frac{C - c_0}{\sum_{h=1}^L \sqrt{c_h} W_h S_h} \frac{W_h S_h}{\sqrt{c_h}}$$

对其求和得到样本量为:

$$n = \frac{(C - c_0) \sum_h \left[\frac{W_h S_h}{\sqrt{c_h}} \right]}{\sum_h W_h S_h \sqrt{c_h}} = \frac{(C - c_0) \sum_h \left[\frac{N_h S_h}{\sqrt{c_h}} \right]}{\sum_h N_h S_h \sqrt{c_h}} \quad (3.29)$$

三、总体参数为 P 的情形

当方差 V 给定时, 如果 N_h 都比较大, 使得

$$N_h - 1 \approx N_h$$

则总样本量为:

(一) 按比例分配

$$n = \frac{\sum W_h P_h Q_h}{V + \frac{\sum W_h P_h Q_h}{N}} \quad (3.30)$$

或
$$n_0 = \frac{\sum W_h P_h Q_h}{V}, n = \frac{n_0}{1 + \frac{n_0}{N}}$$

(二) Neyman 分配

$$n = \frac{\left(\sum W_h \sqrt{P_h Q_h} \right)^2}{V + \frac{\sum W_h P_h Q_h}{N}} \quad (3.31)$$

计算样本量之前, 需要对 P_h 作预估计。

【例 3.5】(续例 3.2) 如果要求在 95% 置信度下, 绝对误差不超过 5%, 则按

比例分配和 Neyman 分配时,总样本量分别为多少?

解:在置信度 95% 时,对应的 $t = 1.96$,而绝对误差 $d = 5\%$,因此

$$V = \left(\frac{d}{t}\right)^2 = \left(\frac{0.05}{1.96}\right)^2 = 0.000651$$

按比例分配时:

由前面的计算结果,可以得到

$$\begin{aligned}\sum W_h P_h Q_h &= \frac{\sum N_h P_h Q_h}{N} \\ &= \frac{1}{2850} (200 \times 0.2 \times 0.8 + 400 \times 0.2 \times 0.8 + 750 \times 0.4 \\ &\quad \times 0.6 + 1500 \times 0.1 \times 0.9) \\ &\approx 0.1442\end{aligned}$$

$$n_0 = \frac{\sum W_h P_h Q_h}{V} = \frac{0.1442}{0.000651} \approx 221.5$$

调整后的样本量为:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{221.5}{1 + \frac{221.5}{2850}} \approx 206$$

Neyman 分配时:

$$\begin{aligned}\sum W_h \sqrt{P_h Q_h} &= \frac{1}{N} \sum N_h \sqrt{P_h Q_h} \\ &= \frac{1}{2850} (200 \times \sqrt{0.2 \times 0.8} + 400 \times \sqrt{0.2 \times 0.8} \\ &\quad + 750 \times \sqrt{0.4 \times 0.6} + 1500 \times \sqrt{0.1 \times 0.9}) \\ &\approx 0.3710\end{aligned}$$

$$\begin{aligned}n &= \frac{(\sum W_h \sqrt{P_h Q_h})^2}{V + \frac{\sum W_h P_h Q_h}{N}} \\ &= \frac{(0.3710)^2}{0.000651 + \frac{0.1442}{2850}} \approx 196\end{aligned}$$

所以,按比例分配和按 Neyman 分配所需的样本量分别为 206 和 196。

§ 3.5 分层时的若干问题

一、抽样效果分析

在实际工作中,通常分层抽样比简单随机抽样的精度要高,也就是说,分层抽样估计量的方差比简单随机抽样的小。由于分层随机抽样的精度与样本量的分配以及各层的方差有关,因此,层的划分或样本量分配不合理时,可能会使分层随机抽样的精度比简单随机抽样的精度还要差。当然,这种情况在理论上可以构造出来,在实际工作中,我们只要不出现不合理地划分层或分配样本量的情况,就可以避免分层随机抽样精度更差的结果发生。

对于固定样本量的情况,如果 $\frac{1}{N_h}$ 相对于 1 可以忽略,则

$$V_{opt} \leq V_{prop} \leq V_{sr} \quad (3.32)$$

式中, V_{opt} , V_{prop} , V_{sr} 分别为分层随机抽样最优分配、分层随机抽样按比例分配以及简单随机抽样简单估计的方差。

如果各层均值差异越大,则采用按比例分配的方式较好,而当各层的标准差相差很大时,则最优分配更好。实际工作中,除非各层的标准差相差很大,人们通常还是喜欢采用按比例分配的方式,这主要是因为最优分配只是针对某个指标(或变量)而言的。实际调查项目中,目标变量通常不止一个,这时,针对某个变量的最优分配,对其他变量可能就是很不合适的,因此,在调查多个目标变量时,按比例分配的分层抽样可能更好些。

对于最优分配,需要各层标准差 S_h 的值,可以用调查指标的历史数据或通过辅助指标的信息推算。也可用与 S_h 有联系的一些量,如层内极差等。

二、层的划分

既然分层抽样比简单随机抽样效率高,那么如何构造层,构造多少层,才能使分层抽样充分发挥其效率高的特点呢?这就涉及最优分层和确定层数的问题。

(一) 最优分层

当分层抽样的使用是为了便于抽样组织、估计子总体的参数,则分层是按自然层或单元的类型划分的。

有时,分层是为了提高抽样效率,这时就要考虑如何进行分层。按调查目标量 Y_i 进行分层当然是最好的,但我们在调查之前并不知道 Y_i 的值,因此分层只能是

通过与 Y_i 高度相关的辅助指标 X_i 来进行。

下面介绍一种确定层界的快速近似法,它是由戴伦纽斯(Dalenius)与霍捷斯(Hodges)提出的。其做法是将分层变量(例如 X_i)分布的累积平方根进行等分来获得最优分层,因此这种方法也称为累积平方根法。下面以一个例子来说明这种方法的实际操作过程。

【例 3.6】 某地区电信部门在对利用电话上网的居民家庭安装 ADSL 意愿进行调查时,以辖区内最近三个月有电话上网支出的居民用户为总体(上网电话费为 0.02 元/分钟),并准备按上网电话费支出(记为 x)进行分层,试确定各层的分点。

表 3.3 前两列给出该市居民家庭上网电话费支出(单位:元)的分布。计算累积频数时应注意, x 区间不是等长的,30 元以下以 5 元为间距,30 元~100 元以 10 元为间距,100 元以上以 50 元为间距,因此计算时,30 元以下的按 \sqrt{f} 累计,30 元~100 元的按 $\sqrt{2f}$ 累计,100 元以上的按 $\sqrt{10f}$ 累计。

表 3.3 居民家庭上网电话费支出分布

范围 x	频数 f	\sqrt{f}	累计 \sqrt{f}
0 ~ 5	65 328	255.593 4	255.593 4
5 ~ 10	89 240	298.730 6	554.324 1
10 ~ 15	36 128	190.073 7	744.397 7
15 ~ 20	77 525	278.433 1	1 022.831
20 ~ 25	62 407	249.813 9	1 272.645
25 ~ 30	24 591	156.815 2	1 429.46
30 ~ 40	24 586	221.747 6	1 651.208
40 ~ 50	9 582	138.434 1	1 789.642
50 ~ 60	15 761	177.544 4	1 967.186
60 ~ 70	8 099	127.271 4	2 094.457
70 ~ 80	5 676	106.545 8	2 201.003
80 ~ 90	3 453	83.102 35	2 284.106
90 ~ 100	4 256	92.260 5	2 376.366
100 ~ 150	1 246	111.624 4	2 487.99
150 ~ 200	800	89.442 72	2 577.433
200 ~ 250	365	60.415 23	2 637.848
250 ~ 300	90	30	2 667.848
300 ~ 350	35	18.708 29	2 686.557
350 ~ 400	5	7.071 068	2 693.628
400 ~ 450	12	10.954 45	2 704.582
> 450	7	8.366 6	2 712.949

最终累计频数是 2 712.949, 如果取层数为 4, 则应每隔 $\frac{2\,712.949}{4} = 678.237$ 分一层, 因此分点应该使得累计 \sqrt{f} 最接近 678.237, 1 356.474, 2 034.712, 即较合理的分层是 $x \leq 15, 15 < x \leq 30, 30 < x \leq 70$ 以及 $x > 70$ (元)。

(二) 层数的确定

当分层是按自然层或单元类型划分时, 层数是自然的, 但当遇到上述运用累积平方根法进行分层时, 就存在确定层数的问题。

在实际工作中, 因为要保证每个层有样本单元, 因此层数不能超过样本量, 如果要给出估计量方差的无偏估计, 则每层至少 2 个样本单元, 那么层数不能超过 $\frac{n}{2}$ 。

通过对分层抽样与简单随机抽样的比较, 我们知道前者比后者的精度高。因此人们设想是否对总体尽可能多地进行划分, 使得层内差异降低, 这时就要涉及层数增加时估计量方差的下降速度。

首先考虑以目标量本身作为分层指标。以最简单的情形为例, Y_i 是区间 d 上的均匀分布, 则总体方差 $S_y^2 = \frac{d^2}{12}$, 样本量为 n 的简单随机抽样简单估计量的方差为 $V(\bar{y}) = \frac{d^2}{12n}$ 。将总体分成大小相同的 L 层, 并按比例分配样本量, 即 $W_h = \frac{1}{L}$, $n_h = \frac{n}{L}$, 则

$$V(\bar{y}_x) = \sum_{h=1}^L W_h^2 \frac{1}{n_h} S_h^2 = \frac{1}{n} \sum_{h=1}^L W_h \frac{d^2}{12L^2} = \frac{d^2}{12nL^2} = \frac{V(\bar{y})}{L^2}$$

由此可见, 层数的增加确实能提高估计精度。

但在工作中, Y_i 本身未知, 只能通过与 Y_i 高度相关的辅助指标 X_i 来进行。这时估计量的方差可以分为两部分, 一部分与层数有关, 另一部分与层数无关, 用模型表示即 $\frac{R^2}{L^2} + (1 - R^2)$, 其中 R^2 是方差中受层数影响的部分, $1 - R^2$ 是不受层数影响的部分。因此, 当层数增加到一定的时候, 在精度上的收益将非常小。根据研究, 除非 Y 与 X 的相关系数 $\rho > 0.95$, 层数一般不超过 6 为宜。

同时, 分层是需要费用的, 因此要考虑增加层数提高的精度与总费用之间的平衡, 因为在总费用一定的条件下, 增加层数必然导致降低样本量, 这时就要考虑增加层数而降低样本量在精度上是否合算。

三、事后分层

对于分层抽样, 我们一般在抽样之前将总体中的所有单元分好层, 但在实际工

作中,有时没有层的抽样框,或总体特别大来不及事先分层,或者几个变量都适合于分层,要进行事先的交叉分层比较困难,并且我们并不需要交叉分层后每个子层的估计,如需要按年龄分层的结果,还需要按受教育程度分层的结果,但并不需要这两个指标的交叉结果。这时如果想利用分层抽样的优点,可以采用对样本的事后分层方法。

要采用事后分层技术,要求我们可以通过某种途径知道各层的层大小 N_h 或层权 W_h 。

事后分层方法还可以用于 y_i 值存在离群值(特别大或特别小)的情况,这时要考虑将总体的离群单元分解,进行事后分层。例如,某市一个样本量为 100 的简单随机样本中,有 15 人最近一年用于购买彩票的支出在 5 000 元以上,我们感觉到这部分人抽多了,对这种极端情况的出现,更改或删除都不太合适,这时最好构造“激进投资者”事后层,并确定总体中这部分人员的真实比例(即层权),通过事后分层对估计结果进行校正。当然,在实际工作中,要得到层权并不容易,这时要决定是利用近似层权进行校正,还是重新抽样。

如果利用事后分层提高估计精度,而层权与实际情况相差很大,则事后分层技术不能达到提高估计精度的目的。例如,利用 10 年前的全国企业普查资料,显然 W_h 变化很大,这时,不能用事后分层技术来对估计进行校正。

使用事后分层技术时,还应注意事后层不宜太多。

最简单的事后分层是先抽取一个样本量为 n 的简单随机样本,然后将样本按某个特征进行分层,落到第 h 层的单元数为 n_h ($\sum_{h=1}^L n_h = n$),则用估计量

$$\hat{y}_{pst} = \sum_{h=1}^L W_h y_h \quad (3.33)$$

来替代样本均值 \bar{y} 。式中, $y_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ 。

当 n_h 固定且都大于零的条件下,落到各层的样本可以看成是独立地从各层中抽取的简单随机样本。这时,事后分层估计量 \bar{y}_{pst} 的方差为:

$$V(\bar{y}_{pst}) = \sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^L W_h S_h^2 \quad (3.34)$$

式中, $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$ 。

理论上,只要 n 充分大,事后分层估计量 y_{pst} 是无偏估计,且它的方差有如下性质:

$$E[V(\bar{y}_{pst})] \approx \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^L (1-W_h) S_h^2 \\ - V_{prop} + \frac{1}{n^2} \sum_{h=1}^L (1-W_h) S_h^2 \quad (3.35)$$

由上式可以看出,第一项就是按比例分配分层抽样估计量的方差,第二项表示因事后分层而非事先按比例分配分层引起的方差增加量。由此看出,只要样本量足够大,事后分层的精度与按比例分配事先分层的精度相当。

如果样本是按某一个辅助指标分层后抽取的,只要这个事先分层抽样是严格按比例分配进行的,则这个样本是自加权的,总体中每个单元被抽中的概率相同,我们可以将这个样本看做简单随机样本,分别对其他指标进行事后分层估计。

【例 3.7】 某高校欲了解在校学生用于课外进修(如各种考证辅导班、外语辅导班等)的开支,在全校 8 000 名学生中抽出了一个 200 人的简单随机样本。根据学生科的统计,本科生人数为全校学生的 70%,调查最近一个学期课外进修支出(单位:元)的结果如表 3.4。

表 3.4 在校学生课外进修开支调查结果

层(h)	层权(W_h)	样本量(n_h)	样本均值(\bar{y}_h)	样本标准差(s_h)
本科生	0.7	120	253.4	231.00
研究生	0.3	80	329.4	367.00
合 计	1	200	283.8	294.57

试估计全校学生用于课外进修的平均开支。

解:全校学生用于课外进修的平均开支为:

$$\bar{y}_{pst} = \sum_{h=1}^2 W_h \bar{y}_h = 0.7 \times 253.4 + 0.3 \times 329.4 = 276.2(\text{元})$$

估计的方差为:

$$v(\bar{y}_{pst}) \approx \frac{1-f}{n} \sum_{h=1}^2 W_h s_h^2 + \frac{1}{n^2} \sum_{h=1}^2 (1-W_h) s_h^2 \\ = \frac{1-0.025}{200} (0.7 \times 231^2 + 0.3 \times 367^2) \\ + \frac{1}{200^2} (0.3 \times 231^2 + 0.7 \times 367^2) \\ = 381.83$$

估计的标准差为:

$$s(y_{\text{pst}}) \approx 19.54(\text{元})$$

如果采用简单估计,则估计的方差为:

$$v(y) = \frac{1-f_s}{n} s^2 = 1 - \frac{0.025}{200} \times 294.57^2 \approx 423.01$$

估计的标准差为:

$$s(y) \approx 20.57(\text{元})$$

小 结

本章介绍了分层抽样理论及若干相关问题。分层抽样技术在实际中应用非常广泛,几乎所有的大型抽样调查项目都要用到分层抽样技术,有时与其他抽样方法结合反复使用。人们之所以喜欢分层抽样技术主要是因为便于项目的组织与管理,同时,其抽样效率通常比简单随机抽样要高。

与简单随机抽样相比,分层抽样在抽样之前需要对总体抽样框进行分层,这个过程有时是现成的,有时需要增加额外的工作量,而且有时可能是相当费时费事的。在推算时需要知道各层的层权或层的大小。

本章附录 分层抽样估计量性质的证明

这里,只给出性质1、2、3的证明,性质4、5、6以及性质7、8、9分别与性质1、2、3对应。

1. 证明性质1:对于一般的分层抽样,如果 \hat{Y}_h 是 Y_h 的无偏估计($h = 1, 2, \dots, L$), 则 \hat{Y}_x 是 \bar{Y} 的无偏估计。

证明:由于对每一层有

$$E(\hat{Y}_h) = \bar{Y}_h$$

因此

$$\begin{aligned} E(\hat{Y}_x) &= E\left(\sum_{h=1}^L W_h \hat{Y}_h\right) = \sum_{h=1}^L W_h E(\hat{Y}_h) \\ &= \sum_{h=1}^L W_h \bar{Y}_h = \frac{1}{N} \sum_{h=1}^L N_h \bar{Y}_h = \frac{1}{N} \sum_{h=1}^L Y_h = \frac{Y}{N} = \bar{Y} \end{aligned}$$

估计量的方差为:

$$V(\hat{Y}_x) = V\left(\sum_{h=1}^L W_h \hat{Y}_h\right) = \sum_{h=1}^L W_h^2 V(\hat{Y}_h) + 2 \sum_{h=1}^L \sum_{k>h}^L W_h W_k \text{Cov}(\hat{Y}_h, \hat{Y}_k)$$

由于各层是独立抽取的,因此上式第二项中的协方差全为零,从而有

$$V(\hat{Y}_x) = \sum_{h=1}^L W_h^2 V(\hat{Y}_h)$$

2. 证明性质 2: 对于分层随机抽样, \bar{y}_x 是 \bar{Y} 的无偏估计。

证明: 对于分层随机抽样, 各层独立进行简单随机抽样, 对每一层有

$$E(y_h) = Y_h$$

因此, 由性质 1, 有

$$E(\bar{y}_x) = \bar{Y}$$

$$V(\bar{y}_x) = \sum_{h=1}^L W_h^2 V(y_h)$$

由第 2 章性质 2, 得

$$V(y_h) = \frac{1}{n_h} f_h S_h^2$$

因此

$$V(\bar{y}_x) = \sum_{h=1}^L W_h^2 V(y_h) = \sum_{h=1}^L W_h^2 \frac{1}{n_h} f_h S_h^2$$

3. 证明性质 3: 对于分层随机抽样, $V(\bar{y}_x)$ 的一个无偏估计为:

$$v(\bar{y}_x) = \sum_{h=1}^L W_h^2 v(\bar{y}_h) = \sum_{h=1}^L W_h^2 \frac{1}{n_h} f_h s_h^2$$

证明: 对于分层随机抽样, 各层独立进行简单随机抽样, 由第 2 章性质 3, 得 $V(y_h)$ 的无偏估计为:

$$v(y_h) = \frac{1}{n_h} f_h s_h^2$$

因此, $V(\bar{y}_x)$ 的一个无偏估计为:

$$v(\bar{y}_x) = \sum_{h=1}^L W_h^2 v(y_h) = \sum_{h=1}^L W_h^2 \frac{1}{n_h} f_h s_h^2$$

4. 在(3.20)条件下, 证明最优分配公式(3.21)。

证明: 对于分层随机抽样, 在线性费用函数条件下, 求最优分配等价于在给定费用 C 时, 选取 n_h 使方差 V 达到最小, 或者在给定方差 V 时, 选取 n_h 使费用 C 达到最小。这个问题等价于极小化下式:

$$V'C' = \left(V + \sum_{h=1}^L \frac{W_h^2}{N_h} S_h^2 \right) (C - c_0) = \sum_{h=1}^L \frac{W_h^2}{n_h} S_h^2 \sum_{h=1}^L c_h n_h$$

式中, V', C' 只包含方差 V 、费用 C 中与样本量有关的部分。

根据柯西—许瓦兹(Cauchy—Schwarz)不等式:

$$(\sum a_h^2)(\sum b_h^2) \geq (\sum a_h b_h)^2$$

等式成立的条件是当且仅当对所有 h , $\frac{b_h}{a_h} = \text{常数}$ 。取

$$a_h = \frac{W_h S_h}{\sqrt{n_h}}, b_h = \sqrt{c_h n_h}$$

于是,当

$$\frac{b_h}{a_h} = \frac{\sqrt{c_h n_h}}{\frac{W_h S_h}{\sqrt{n_h}}} = \frac{n_h \sqrt{c_h}}{W_h S_h} = K = \text{常数}$$

也即

$$n_h = K \frac{W_h S_h}{\sqrt{c_h}}$$

对所有 h 成立时, $V'C'$ 达到极小。

对所有 h 求和,有

$$n = \sum_{h=1}^L n_h = K \sum_{h=1}^L \frac{W_h S_h}{\sqrt{c_h}}$$

因此,最优分配为:

$$\frac{n_h}{n} = \frac{\frac{W_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^L \frac{W_h S_h}{\sqrt{c_h}}} = \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^L \frac{N_h S_h}{\sqrt{c_h}}}$$

5. 证明式(3.32)。

证明:由最优分配的定义, $V_{opt} \leq V_{prop}$ 。因此下面只需证明 $V_{prop} \leq V_{srs}$, 即只要证明

$$V_{prop} = \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2 \leq \frac{1-f}{n} S^2 = V_{srs}$$

成立即可。

对总体离差平方和进行分解,得

$$(N-1)S^2 = \sum_{h=1}^L \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y})^2$$

$$= \sum_{h=1}^L \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 + \sum_{h=1}^L N_h (Y_h - \bar{Y})^2$$

$$= \sum_{h=1}^L (N_h - 1) S_h^2 + \sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2$$

等式两边同除以 $N - 1$, 若对所有的 h , $\frac{1}{N_h}$ 相对于 1 可以忽略, 则有

$$\frac{N_h}{N - 1} \approx \frac{N_h - 1}{N - 1} \approx W_h$$

于是

$$S^2 \approx \sum_{h=1}^L W_h S_h^2 + \sum_{h=1}^L W_h (\bar{Y}_h - \bar{Y})^2$$

注意上式中第二项是非负的, 因此

$$V_{srs} = \frac{1}{n} f S^2 \approx \frac{1}{n} f \sum_{h=1}^L W_h S_h^2 + \frac{1}{n} f \sum_{h=1}^L W_h (\bar{Y}_h - \bar{Y})^2$$

$$= V_{prop} + \frac{1-f}{n} \sum_{h=1}^L W_h (\bar{Y}_h - \bar{Y})^2$$

从而有

$$V_{opt} \leq V_{prop} \leq V_{srs}$$

习 题

1. 某高校欲了解教职员工对某项津贴与职务职称挂钩的分配制度改革的态度, 准备在全校教职员工中进行抽样调查。为了提高抽样效率, 准备进行分层抽样, 请判断下面的几种分层方法是否合适:

- (1) 按性别分层;
- (2) 按教师、行政管理人员、职工分层;
- (3) 按职称(正高、副高、中级、初级、其他)分层;
- (4) 按部门(如系、所、处)分层。

2. 某学院 4 个专业的新生举行元旦晚会, 组织者为了活跃气氛, 欲在 200 名学生中抽出 10 名作为“幸运星”, 为了以示公平, 要求每位学生被抽中的概率相同。组织者知道利用简单随机抽样的方法可以满足要求, 你能否帮助组织者再设计几种方案?

3. 某居委会辖有三个居民新村, 居委会欲对居民购买彩票的情况进行调查。

调查者考虑以新村分层,在每个新村中随机抽取了 10 个居民户并调查每户最近一个月购买彩票花费的金额(元),下表是每个新村及调查的情况:

新 村	居民户数	1	2	3	4	5	6	7	8	9	10
1	256	10	10	2	0	20	10	0	10	30	20
2	420	20	35	10	50	0	40	50	10	20	20
3	168	0	20	0	30	30	50	40	0	30	0

(1) 试估计该小区居民户购买彩票的平均支出,并给出估计的标准差;

(2) 当置信度为 95%,要求相对误差不超过 10% 时,按比例分配和 Neyman 分配时样本量及各层的样本量分别为多少?

4. 随着经济发展,某市居民正在悄悄改变过年的习惯,虽然仍有大多数居民除夕夜在家吃年夜饭、看电视节目,但也有些家庭到饭店吃年夜饭,或逛夜市,或利用过年的假期到外地旅游。为研究这种现象,某研究机构以市中心 165 万居民户作为研究对象,将居民户按 6 个行政区分层,在每个行政区随机抽出 30 户居民户进行了调查(各层抽样比可以忽略),每个行政区的情况以及在家吃年夜饭、看电视节目的居民户比例如下表:

行政区(h)	居民户比例(W_h)	在家居民户(n_h)
1	0.18	27
2	0.21	28
3	0.14	27
4	0.09	26
5	0.16	28
6	0.22	29

(1) 试估计该市居民在家吃年夜饭的比例,并给出估计的标准差;

(2) 当置信度为 95%,要求绝对误差不超过 1% 时,按比例分配和 Neyman 分配时总样本量及各层的样本量分别为多少?

5. 某开发区利用电话调查(RDD)对区内居民消费冷冻食品情况进行调查,他们将电话号码(六位数字)的前两位作为一部分,后四位作为一部分,前两位代表局号,局号及每个局号中拥有的电话数可以找到,按局号分层,按每个局号(剔除商户后)拥有的电话数比例分配样本量(各层抽样比可以忽略)。调查后各层样本户

购买冷冻食品支出的中间结果如下表：

局号	层权(%)	样本量	样本平均(元)	样本标准差
1	8.2	16	89	105
2	6.5	13	56	74
3	13.7	27	102	186
4	5.6	11	76	97
5	11.8	24	97	106
6	11.6	23	79	89
7	17.0	34	83	112
8	9.8	20	52	73
9	8.8	18	36	44
10	7.0	14	52	65

试估计该开发区居民户购买冷冻食品的平均支出,以及估计的 95% 置信区间。

6. 某单位欲估计职工的离职意愿,聘请了专业公司来进行调研,公司人员按高级职称、中级职称和初级职称分为三层,已知层权分别为 0.2,0.3,0.5,预先猜测各层的总体比例为 0.1,0.2,0.4,如果采用按比例分配的分层抽样,要求估计的方差与样本量为 100 的简单随机样本相当,则样本量应为多少(不考虑有限总体校正系数)?

7. 如果一个大的简单随机样本,按类别分为 6 组,然后按照层的实际大小重新进行加权,这一过程称为事后分层,采用这种方法是由于(判断以下说法的对错):

- (1) 它能比简单随机抽样产生更精确的结果;
- (2) 它能比按比例分配产生更精确的结果;
- (3) 它能比最优分配产生更精确的结果;
- (4) 在抽样时不能得到分层变量;
- (5) 它的估计量的方差与真正按比例分层随机抽样的方差差不多。

8. 某公司进行财务审计,需要对原始凭证进行审核,该公司先后有两名出纳,由 A 出纳登记的原始凭证占 70%,B 出纳登记的原始凭证占 30%。审计人员从原始凭证中随机抽出 100 份,结果发现,由 A,B 出纳登记的原始凭证分别为 43 份和 57 份,差错分别为 1 份和 2 份。

(1) 用简单随机抽样的公式估计登记原始凭证的差错率,并计算估计的标准差;

(2) 用事后分层的公式估计登记原始凭证的差错率,并计算估计的标准差(有限总体校正系数 $1 - f \approx 1$)。



第 4 章

比率、回归与差值估计

调查时需要推算的目标量分为总体总量、均值、比例及比率。前面介绍了对总体总量、均值以及比例的简单估计,简单估计是线性的。比如对于总体均值,在简单随机抽样时,用样本均值进行估计;在分层抽样时,用各层样本均值的加权平均来估计。对总体比率的估计不同于前三种目标量,它需要用非线性估计,这就是本章将介绍的比率估计量。

在实际工作中,如果除了调查的目标量以外,还有其他指标的信息,称这些指标为辅助变量(auxiliary variable)。人们总希望利用这些辅助变量与目标量之间的关系提高估计精度,这时,可以考虑利用本章介绍的几种估计方法。

§ 4.1 引言

一、概念与作用

(一) 概念

当调查的目标量是总体比率时,所用的估计量不同于对比例的估计。因为前者涉及总体两个指标,这两个指标都需要通过样本进行估计,而后者涉及的总体大小

是已知的,不需要估计。例如,在对全国货物运输量进行统计时,目标量为全国总货运量、总货物周转量,由这两个量可以得到货物的平均运输距离,称为平均运距。即

$$\text{平均运距} = \frac{\text{总货物周转量}}{\text{总货运量}}$$

由于全国总货运量、总货物周转量都需要通过样本进行估计,因此平均运距本身是一个比率量。又如,家庭用于教育的支出占总支出的比重,家庭教育支出以及总支出都需要估计。再如,拨号上网的网民家庭中安装 ISDN 的比重。

对总体进行调查时,调查的指标往往是多个,除了调查指标之外,还有其他指标(辅助变量),这时人们考虑利用其他指标的信息来提高调查指标估计的精度。通常是利用调查指标与辅助变量之间的关系构造比率估计量或回归估计量。

(二) 作用

在进行抽样调查时,目标量本身就是总体比率,这时,对总体比率的估计要用到本章介绍的比率估计量。大多数情况下,人们利用比率估计、回归估计,都是希望利用总体的辅助信息来提高估计的精度。通常,只要调查指标与辅助变量存在较好的正相关关系,比率估计、回归估计就比简单估计好。

比率估计、回归估计同样也可以用于分层随机抽样,而且分层比率估计、分层回归估计比通常的分层简单估计要好。

二、应用条件

比率估计、回归估计是非线性估计,与简单估计相比,其优劣取决于辅助变量的选择,也就是辅助变量应该与调查指标有较好的正相关关系,例如成比例关系或线性回归关系。

如果辅助变量与调查指标具有较好的负相关关系,则要采用乘积估计。由于实际工作中具有负相关关系的辅助变量的情形很少见,因此,理论上给出了乘积估计的公式,但实际案例很少见到。

比率估计、回归估计需要用到辅助变量的总体均值,因此辅助变量的总体总量或总体均值应该是已知的。实际工作中,如果辅助变量的总体总量或总体均值未知,又要利用比率估计或回归估计,则可以采用二重抽样方法,先获得辅助变量的估计,再对目标量进行估计。

比率估计是有偏估计,回归估计中如果用样本回归系数时,回归估计也是有偏估计。但当样本量足够大时,估计的偏倚趋于零,因此,比率估计、回归估计需要有足够的样本量才能保证估计的有效。

三、符号说明

设调查指标为 Y , 辅助变量为 X 。本章将用到目标变量和辅助变量的如下指标:

$$\text{总体总量: } Y = \sum_{i=1}^N Y_i, X = \sum_{i=1}^N X_i$$

$$\text{总体均值: } Y = \frac{1}{N} \sum_{i=1}^N Y_i, X = \frac{1}{N} \sum_{i=1}^N X_i$$

$$\text{总体方差: } S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - Y)^2, S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

$$\text{样本均值: } y = \frac{1}{n} \sum_{i=1}^n y_i, x = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\text{样本方差: } s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - y)^2, s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{总体协方差: } S_{xy} = \frac{1}{N-1} \sum_{i=1}^N (Y_i - Y)(X_i - \bar{X})$$

$$\text{样本协方差: } s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (y_i - y)(x_i - \bar{x})$$

$$\text{总体相关系数: } \rho = \frac{S_{xy}}{S_x S_y}$$

$$\text{样本相关系数: } \hat{\rho} = \frac{s_{xy}}{s_x s_y}$$

§ 4.2 比率估计

一、简单随机抽样下的比率估计

(一) 定义

比率估计量(ratio estimator) 又称比估计。对于简单随机抽样, 总体均值 Y 和总体总量 Y 的比率估计为:

$$y_R = \frac{y}{x} \bar{X} = \frac{\sum y_i}{\sum x_i} X \quad (4.1)$$

$$\hat{Y}_R = \frac{y}{x} X = \frac{\sum y_i}{\sum x_i} X = N \bar{y}_R \quad (4.2)$$

有时, 调查的目标量就是总体比率:

$$R = \frac{Y}{X} = \frac{\sum_{i=1}^N Y_i}{\sum_{i=1}^N X_i} \quad (4.3)$$

对总体比率的估计为样本比率:

$$\hat{R} = \frac{y}{x} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} \quad (4.4)$$

(二) 比率估计的性质

简单随机抽样比率估计是有偏的,其偏倚的阶为 $O\left(\frac{1}{n}\right)$,因此当样本量 n 较大时,估计量的偏倚趋于零,因此,比率估计是渐近无偏的.

性质 1 对于简单随机抽样比率估计,当样本量 n 较大时, y_R , \hat{Y}_R 及 \hat{R} 是渐近无偏的,即

$$E(\bar{y}_R) \approx Y, E(\hat{Y}_R) \approx Y, E(\hat{R}) \approx R \quad (4.5)$$

y_R , \hat{Y}_R 及 \hat{R} 的方差为:

$$\begin{aligned} V(y_R) &\approx \frac{1}{n} \frac{f}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 \\ &= \frac{1}{n} \frac{f}{N} (S_y^2 + R^2 S_x^2 - 2RS_{yx}) \end{aligned} \quad (4.6)$$

$$\begin{aligned} V(\hat{Y}_R) &\approx \frac{N^2(1-f)}{n} \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 \\ &= \frac{N^2(1-f)}{n} (S_y^2 + R^2 S_x^2 - 2RS_{yx}) \end{aligned} \quad (4.7)$$

$$\begin{aligned} V(\hat{R}) &\approx \frac{1-f}{nX^2} \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 - \frac{1-f}{nX^2} (S_y^2 + R^2 S_x^2 - 2RS_{yx}) \end{aligned} \quad (4.8)$$

$v(\hat{r})$ 的样本估计式为:

$$v_1(\hat{R}) \approx \frac{1}{nX^2} f (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}) \quad (4.9)$$

$$\text{或} \quad v_2(\hat{R}) \approx \frac{1-f}{nX^2} (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}) \quad (4.10)$$

式中 S_y^2, S_x^2, S_{yx} 分别为 Y, X 的总体方差和总体协方差; s_y^2, s_x^2, s_{yx} 分别为 Y, X 的样本方差和样本协方差

$V(y_R) = X^2 V(\hat{R})$, $V(\hat{Y}_R) = X^2 V(\hat{R})$ 可通过 $v_1(\hat{R})$ 或 $v_2(\hat{R})$ 估计

【例 4.1】对以下假设总体 ($N = 6$), 用简单随机抽样抽取 $n = 2$ 的样本 (见表 4.1), 比较简单随机抽样比率估计及简单估计的性质。

表 4.1 假设的总体数据

i	1	2	3	4	5	6	均值
X_i	0	1	3	5	8	10	4.5
Y_i	1	3	11	18	29	46	18

解: 对这个总体, 我们列出所有可能的 $C_6^2 = 15$ 个样本, 以比较简单估计和比率估计的性质。

i	样本	简单估计(\bar{y})	比率估计(y_R)
1	1, 2	2.0	18
2	1, 3	6.0	18
3	1, 4	9.5	17.1
4	1, 5	15.0	16.875
5	1, 6	23.5	21.15
6	2, 3	7.0	15.75
7	2, 4	10.5	15.75
8	2, 5	16.0	16
9	2, 6	24.5	20.045 5
10	3, 4	14.5	16.312 5
11	3, 5	20.0	16.363 6
12	3, 6	28.5	19.730 8
13	4, 5	23.5	16.269 2
14	4, 6	32.0	19.2
15	5, 6	37.5	18.75

由此, 可以计算出:

$$E(y) = \frac{1}{15} \sum_{i=1}^{15} y_i = \frac{2 + 6 + \cdots + 37.5}{15} = 18$$

$$V(y) = \frac{1}{15} \sum_{i=1}^{15} [y_i - E(y)]^2 \approx 97.866 67$$

$$E(y_R) = \frac{1}{15} \sum_{i=1}^{15} \bar{y}_{Ri} = \frac{18 + 18 + \cdots + 18.75}{15} \approx 17.686 44$$

$$B(y_R) = E(y_R) - Y \approx 17.686 44 - 18 = -0.313 56$$

$$V(y_R) = \frac{1}{15} \sum_{i=1}^{15} [y_{Ri} - E(y_R)]^2 \approx 2.82345$$

$$MSE(y_R) = V(y_R) + B^2(y_R) \approx 2.82345 + (-0.31356)^2 = 2.92177$$

由计算结果可以看出,简单估计是无偏的,而比率估计是有偏的。简单估计量的方差远远大于比率估计量的方差,比率估计的偏倚不大,其均方误差也比简单估计的小得多。因此,对这个总体,比率估计比简单估计的效率高。

【例 4.2】 某县在对船舶调查月完成的货运量进行调查时,对运管部门登记的船舶台账进行整理后获得注册船舶 2 860 艘,载重吨位 154 626 吨。从 2 860 艘船舶中抽取了一个 $n = 10$ 的简单随机样本,调查得到样本船舶调查月完成的货运量及其载重吨位如表 4.2(单位:吨),要推算该县船舶调查月完成的货运量。

表 4.2 样本船舶货运量及载重吨位数据

i	y_i	x_i	i	y_i	x_i
1	780	100	6	2 170	120
2	1 500	50	7	1 823	150
3	1 005	50	8	1 450	80
4	376	10	9	158	20
5	600	20	10	1 370	50

解:已知: $N = 2\,860$, $n = 10$, $X = 154\,626$

由表 4.2 可得

$$\bar{y} = \frac{1}{10} \sum_{i=1}^{10} y_i = 1\,123.2, \bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 65$$

$$s_y^2 = \frac{1}{10-1} \sum_{i=1}^{10} (y_i - \bar{y})^2 \approx 421\,179.07$$

$$s_x^2 = \frac{1}{10-1} \sum_{i=1}^{10} (x_i - \bar{x})^2 \approx 2\,161.11$$

$$s_{xy} = \frac{1}{10-1} \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) \approx 23\,382.22$$

因此,对该县船舶在调查月完成货运量的比率估计为:

$$\hat{Y}_R = \frac{y}{x} X = \frac{1\,123.2}{65} \times 154\,626 = 2\,671\,937(\text{吨})$$

\hat{Y}_R 方差的估计为:

$$v(\hat{Y}_R) \approx \frac{N^2(1-f)}{n} (s_y^2 + \hat{R}^2 s_x^2 - 2 \hat{R} s_{xy}) = 2.10617 \times 10^{11}$$

\hat{Y}_R 标准差的估计为:

$$s(\hat{Y}_R) = \sqrt{v(\hat{Y}_R)} \approx 458930(\text{吨})$$

如果用简单估计对货运量进行估计,则

$$\hat{Y} = N\bar{y} = 2860 \times 1123.2 = 3212352(\text{吨})$$

$$v(\hat{Y}) = \frac{N^2(1-f)}{n} s_y^2 = 3.43303 \times 10^{11}$$

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} \approx 585921(\text{吨})$$

由此,得到比率估计量设计效应的估计为:

$$deff = \frac{v(\hat{Y}_R)}{v(\hat{Y})} \approx 0.6135$$

对于本问题,比率估计量的效率比简单估计量的效率高。注意,这里只是作为一个例子,实际中对于样本量较小的情形,使用比率估计量时不能忽视其偏倚。

(三) 消除比率估计偏倚的方法

由于比率估计是有偏估计,在小样本时,其偏倚不能忽略。如果这时有很好的辅助变量,希望使用比率估计来提高精度,则需要通过改善估计量或改变抽样方法使比率估计成为无偏估计。

1. 无偏的比率型估计量。这里主要介绍两种无偏的比率型估计量。

第一种无偏的比率型估计量是哈特利-罗斯(Hartley-Ross)估计量。它从比率 $\frac{y_i}{x_i}$ 的平均值 \bar{r} 出发,然后校正 r 的偏倚获得。哈特利-罗斯估计量为:

$$\hat{R}_{HR} = r + \frac{n(N-1)}{(n-1)X} (y - r \cdot x) \quad (4.11)$$

式中,

$$r = \frac{1}{n} \sum_{i=1}^n r_i = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i} \quad (4.12)$$

等式右边第二项是对 \bar{r} 的偏倚 $E(r) - R$ 的修正。

第二种无偏的比率型估计量是米基(Mickey)估计量。与哈特利-罗斯估计量类似,它也是从比率的平均值出发,但这时用 $\hat{R}_{-i} = \frac{y_{-i}}{x_{-i}}$ 代替上面的 r_i 。这里 y_{-i} , x_{-i} 分别表示在 n 个样本数据中去掉第 i 个样本数据后剩下的 $n-1$ 个样本数据的

平均值。记 \hat{R}_M 的平均值为 R ，米基估计量为：

$$\hat{R}_M = R + \frac{n(N-n+1)}{X}(y - R \cdot x) \quad (4.13)$$

【例 4.3】对如下一个 $N = 5$ 的假设总体，总体比率为 $R = 2$ 。如果样本量 $n = 3$ ，说明哈特利—罗斯估计量和米基估计量的无偏性(见表 4.3)。

表 4.3 假设的总体数据

i	1	2	3	4	5	平均值
Y_i	2	4	5	9	10	6
X_i	1	2	3	4	5	3

解：将 $n = 3$ 的所有可能样本列于表 4.4，并计算每个样本的估计量(结果中只列出小数点后三位)。

表 4.4

样本	y	x	\hat{R}	r	\hat{R}_{HR}	R	\hat{R}_M
1,2,3	3.667	2.000	1.833	1.889	1.844	1.850	1.830
1,2,4	5.000	2.333	2.143	2.083	2.139	2.122	2.151
1,2,5	5.333	2.667	2.000	2.000	2.000	2.000	2.000
1,3,4	5.333	2.667	2.000	1.972	2.002	1.983	2.010
1,3,5	5.667	3.000	1.889	1.889	1.889	1.875	1.900
1,4,5	7.000	3.333	2.100	2.083	2.106	2.104	2.096
2,3,4	6.000	3.000	2.000	1.972	2.006	1.989	2.009
2,3,5	6.333	3.333	1.900	1.889	1.904	1.892	1.908
2,4,5	7.667	3.667	2.091	2.083	2.094	2.093	2.089
3,4,5	8.000	4.000	2.000	1.972	2.017	1.995	2.006
平均值	6.000	3.000	1.996	1.983	2.000	1.990	2.000

从上述计算中，可以看出， \hat{R} 是有偏的，而 \hat{R}_{HR} 和 \hat{R}_M 是无偏的。

2. 改变抽样方法。使比率估计成为无偏估计的另一种办法是改变抽样方法。

拉希里(Lahiri)证明,只要每个大小为 n 的样本被抽中的概率与其辅助变量的和

$\sum_{i=1}^n x_i$ 成比例,则这时的比率估计就是无偏估计。

为获得满足这个条件的样本,最简单的办法可能是水野(Midzuno)法,即在总体中按与 X_i 成比例的概率抽取第一个样本单元,在总体剩下的单元中按简单随机抽样抽取 $n-1$ 个样本单元,则这 n 个单元组成的样本被抽中的概率与 $\sum_{i=1}^N x_i$ 成比例。

二、分层随机抽样下的比率估计

对于分层随机抽样,如果采用比率估计量,由于比率估计量是有偏的,只有在大量样本的条件下,偏倚才趋于零,因此如果各层的样本量比较大,则可以采用各层分别进行比率估计,将各层加权汇总得到总体指标的估计,这种方式称为分别比率估计。

有时各层只是一个小样本,使用分别比率估计可能效果不好,这时可以采用联合比率估计。

(一) 分别比率估计

总体均值 \bar{Y} 和总体总量 Y 的分别比率估计量(separate ratio estimator)为:

$$y_{Rs} = \sum_{h=1}^L W_h y_{Rh} = \sum_{h=1}^L W_h \frac{y_h}{x_h} X_h \quad (4.14)$$

$$\hat{Y}_{Rs} = N y_{Rs} = \sum_{h=1}^L \frac{y_h}{x_h} X_h = \sum_{h=1}^L \hat{Y}_{Rh} \quad (4.15)$$

式中, W_h 为层权; L 为层数; y_h 和 x_h 分别为 Y_h 和 X_h 的简单估计; y_{Rh} 和 \hat{Y}_{Rh} 分别为 Y_h 和 Y_h 的比率估计。

如果每一层的样本量 n_h 较大,则每一层的比率估计是近似无偏的,因此这时分层随机抽样分别比率估计量也是近似无偏的,并且由每一层比率估计量的方差得到分别比率估计量的方差:

$$V(y_{Rs}) \approx \sum_{h=1}^L \frac{W_h^2 (1 - f_h)}{n_h} (S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h \rho_h S_{yh} S_{xh}) \quad (4.16)$$

$$V(\hat{Y}_{Rs}) \approx \sum_{h=1}^L \frac{N_h^2 (1 - f_h)}{n_h} (S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h \rho_h S_{yh} S_{xh}) \quad (4.17)$$

式中, $f_h = \frac{n_h}{N_h}$; $S_{yh}^2, S_{xh}^2, \rho_h$ 分别为第 h 层指标 Y, X 的方差及其相关系数。

分别比率估计量要求每一层的样本量都比较大,如果达不到这个要求,则它的偏倚可能比较大,这时使用联合比率估计量可能更好些。

(二) 联合比率估计

总体均值 \bar{Y} 和总体总量 Y 的联合比率估计量(combined ratio estimator)为:

$$y_{Rc} = \frac{y_s}{x_s} X - \hat{R}_c X \quad (4.18)$$

$$\hat{Y}_{Rc} = \frac{y_s}{x_s} X - \hat{R}_c X \quad (4.19)$$

式中, y_s 和 x_s 分别为 Y 和 X 的分层估计。

分层随机抽样联合比率估计量是有偏的,但当总样本量 n 较大时,估计量的偏倚趋于零,因此,联合比率估计量是渐近无偏的。即

$$E(\bar{y}_{Rc}) \approx \bar{Y}, E(\hat{Y}_{Rc}) \approx Y \quad (4.20)$$

\bar{y}_{Rc} , \hat{Y}_{Rc} 的均方误差为:

$$MSE(\bar{y}_{Rc}) \approx V(\bar{y}_{Rc}) \approx \sum_h \frac{N_h^2(1 - \frac{f_h}{N})}{N^2 n_h} (S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{yxh}) \quad (4.21)$$

$$MSE(\hat{Y}_{Rc}) \approx V(\hat{Y}_{Rc}) \approx \sum_h \frac{N_h^2(1 - \frac{f_h}{N})}{n_h} (S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{yxh}) \quad (4.22)$$

将各指标的样本估计代入就可得到均方误差的样本估计。

(三) 分别比率估计量与联合比率估计量的比较

如果每一层都满足比率估计量有效的条件,则除非 $R_h = R$,都有分别比率估计量的方差小于联合比率估计量的方差。但当每层的样本量不太大时,还是采用联合比率估计量更可靠些,因为这时分别比率估计量的偏倚很大,从而使总的均方误差增大。

实际使用时,如果各层的样本量都较大,且有理由认为各层的比率 R_h 差异较大,则分别比率估计优于联合比率估计。当各层的样本量不大,或各层比率 R_h 差异很小,则联合比率估计更好些。

【例 4.4】 某市 1996 年对 950 家港口生产单位完成的吞吐量进行了调查,1997 年欲对全市港口生产单位完成的吞吐量进行抽样调查。对港口生产单位按非国有($h = 1$)和国有($h = 2$)分为两层,单位数分别为 800 家和 150 家,分别在两层中调查了 10 家、15 家港口生产单位,调查数据如表 4.5,试估计 1997 年全市港口生

产单位完成的吞吐量。

表 4.5 1997 年非国有和国有企业调查数据

i	x_i	y_i	i	x_i	y_i
1	95	80	1	495	530
2	220	210	2	210	320
3	359	384	3	360	496
4	120	117	4	230	400
5	177	180	5	600	651
6	253	258	6	1 000	880
7	302	349	7	700	560
8	332	286	8	1 100	1 230
9	272	215	9	720	823
10	137	97	10	310	390
			11	478	465
			12	817	650
			13	919	1 160
			14	1 160	1 070
			15	735	698

解:将上述数据计算的中间结果列于表 4.6。

表 4.6

	h 1. 非国有	h 2. 国有	合 计
n_h	10	15	25
N_h	800	150	950
W_h	0.842 105	0.157 895	1
f_h	0.012 5	0.1	
X_h	171 400	102 900	274 300
\bar{X}_h	214.25	686	
\bar{x}_h	226.7	655.6	
\bar{y}_h	217.6	688.2	
s_{mh}^2	8 477.344	94 665.26	
s_{ph}^2	10 704.71	82 541.89	
s_{mth}	9 072.2	81 071.51	
\hat{R}_h	0.959 859	1.049 725	

1. 按分别比率估计量估计。

$$\begin{aligned}\hat{Y}_{RS} &= \sum_{h=1}^2 \hat{R}_h X_h = 0.959\ 859 \times 171\ 400 + 1.049\ 725 \times 102\ 900 \\ &\approx 272\ 536.5\end{aligned}$$

$$\begin{aligned}v(\hat{Y}_{RS}) &= \sum_{h=1}^2 \frac{N_h^2(1-f_h)}{n_h} (s_{yh}^2 + \hat{R}_h^2 s_{xh}^2 - 2\hat{R}_h s_{yxh}) \\ &= 69\ 461\ 324.15 + 22\ 477\ 628.53 = 91\ 938\ 952.68\end{aligned}$$

$$s(\hat{Y}_{RS}) = \sqrt{v(\hat{Y}_{RS})} = 9\ 588.48$$

2. 按联合比率估计量估计。

$$\hat{Y}_s = \sum_{h=1}^2 N_h y_h = 800 \times 217.6 + 150 \times 688.2 = 277\ 310$$

$$\hat{X}_s = \sum_{h=1}^2 N_h x_h = 800 \times 226.7 + 150 \times 655.6 = 279\ 700$$

$$\hat{Y}_{RC} = \frac{\hat{Y}_s}{\hat{X}_s} X = \frac{277\ 310}{279\ 700} \times 274\ 300 = 271\ 956.1$$

$$\begin{aligned}v(\hat{Y}_{RC}) &= \sum_{h=1}^2 \frac{N_h^2(1-f_h)}{n_h} (s_{yh}^2 + \hat{R}^2 s_{xh}^2 - 2\hat{R} s_{yxh}) \\ &= 66\ 261\ 436.65 + 20\ 032\ 262.19 = 86\ 293\ 698.84\end{aligned}$$

$$s(\hat{Y}_{RC}) = \sqrt{v(\hat{Y}_{RC})} = 9\ 289.44$$

三、比率估计的效率

(一) 与简单估计的比较

对于简单随机抽样,简单估计量是无偏的,而比率估计量是渐近无偏的,因此这里只比较当 n 比较大的情形。为了比较简单估计和比率估计的优劣,可通过比较它们的均方误差或方差大小来进行。

由上面的讨论,我们知道:

$$V(\bar{y}) = \frac{1-f}{n} S_y^2 \quad (4.23)$$

$$\begin{aligned}V(y_R) &\approx \frac{1-f}{n} (s_v^2 + R^2 S_x^2 - 2R S_{vx}) \\ &= \frac{1-f}{n} (S_v^2 + R^2 S_x^2 - 2R \rho S_y S_x) \quad (4.24)\end{aligned}$$

由此可以看出,比率估计量优于简单估计量的条件是

$$R^2 S_x^2 - 2R\rho S_y S_x < 0$$

整理后,得到当

$$\rho > \frac{1}{2} \frac{\frac{S_x}{\bar{Y}}}{\frac{S_y}{\bar{Y}}} \frac{C_x}{2C_y} \quad (4.25)$$

有 $V(y_R) < V(y)$

特别当 $C_x \approx C_y$ 时, $\rho > \frac{1}{2}$, 比率估计量就优于简单估计量。

(二) 比率估计成为最优线性估计的条件

当总体满足下面两个条件时,则比率估计是最优线性估计:(1) y_i 与 x_i 的关系是过原点的直线;(2) y_i 对这条直线的方差与 x_i 成比例。

§ 4.3 回归估计

类似比率估计量,如果除了调查指标(Y)之外,还有其他指标(X)可利用, X 称为辅助变量, Y 与 X 有较好的相关关系,且 Y 对 X 的回归线不通过原点,则可利用调查指标与辅助变量之间的相关关系来提高估计的精度。但是 X 的总体总量或总体均值应该是已知的

一、回归估计的定义

对于简单随机抽样,总体均值 Y 和总体总量 Y 的回归估计量(regression estimator)的定义式为:

$$y_{lr} = \bar{y} + \beta(X - \bar{x}) = \bar{y} - \beta(\bar{x} - X) \quad (4.26)$$

$$\hat{Y} = N\hat{y}_{lr} \quad (4.27)$$

式中, \bar{y}, \bar{x} 为样本均值; β 为事先设定的一个常数,也可以由样本决定,例如样本回归系数。

如果 $\beta = 0$,则回归估计量就是简单估计量;如果 $\beta = \frac{\bar{y}}{\bar{x}}$,则回归估计量就是比率估计量

二、 β 为常数的情况

当回归系数 β 为事先给定的常数时,或以前为相同目的进行的调查所得到的 Y_i 对 X_i 的样本回归系数 $\hat{\beta}$ 稳定在某个数值上,取最近一次调查所得的 $\hat{\beta}$ 作为设定值

性质 2 对于简单随机抽样回归估计量,作为 Y 及 Y 的回归估计, y_{lr} 及 \hat{Y}_{lr} 都是无偏的。即

$$E(y_{lr}) = Y$$

$$E(\hat{Y}_{lr}) = E(Ny_{lr}) = Y \quad (4.28)$$

y_{lr} 和 \hat{Y}_{lr} 的方差为:

$$V(y_{lr}) = \frac{1-f}{n} (S_y^2 + \beta_0^2 S_x^2 - 2\beta_0 S_{yx}) \quad (4.29)$$

$$V(\hat{Y}_{lr}) = \frac{N^2(1-f)}{n} (S_y^2 + \beta_0^2 S_x^2 - 2\beta_0 S_{yx}) \quad (4.30)$$

$V(y_{lr})$ 和 $V(\hat{Y}_{lr})$ 的样本估计为:

$$v(y_{lr}) = \frac{1-f}{n} (s_y^2 + \beta_0^2 s_x^2 - 2\beta_0 s_{yx}) \quad (4.31)$$

$$v(\hat{Y}_{lr}) = \frac{N^2(1-f)}{n} (s_y^2 + \beta_0^2 s_x^2 - 2\beta_0 s_{yx}) \quad (4.32)$$

式中, S_y^2, S_x^2, S_{yx} 分别为 Y, X 的总体方差和总体协方差; s_y^2, s_x^2, s_{yx} 分别为 Y, X 的样本方差和样本协方差。

当 β_0 取总体回归系数

$$B = \frac{S_{yx}}{S_x^2} = \frac{\sum_{i=1}^N (Y_i - Y)(X_i - X)}{\sum_{i=1}^N (X_i - \bar{X})^2} \quad (4.33)$$

时, $V(y_{lr})$ 达到最小, 即

$$V_{\min}(y_{lr}) = \frac{1-f}{n} (S_y^2 - B^2 S_x^2) = \frac{1-f}{n} S_y^2 (1 - \rho^2) \quad (4.34)$$

式中, ρ 为 Y_i 与 X_i 总体相关系数。

三、 β 为样本回归系数的情况

如果 β 需要通过样本来确定,很自然地,我们会想到用总体回归系数的最小二乘估计,也就是样本回归系数:

$$b = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.35)$$

这时简单随机抽样回归估计量

$$y_{lr} = \bar{y} + b(X - \bar{x}) \quad (4.36)$$

是有偏的,但当样本量 n 充分大时,估计量的偏倚趋于零,因此,类似比率估计量,回归估计量也是渐近无偏的,且

$$MSE(\bar{y}_{lr}) \approx V(\bar{y}_{lr}) \approx \frac{1-f}{n} S_y^2 (1 - \rho^2) \quad (4.37)$$

$MSE(\bar{y}_{lr})$ 和 $V(\bar{y}_{lr})$ 的一个近似估计为:

$$v(\bar{y}_{lr}) = \frac{1-f}{n} s_e^2 = \frac{1-f}{n(n-2)} \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \quad (4.38)$$

式中, $s_e^2 = \frac{1}{n-2} \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$

$$= \frac{1}{n-2} \left[\sum_{i=1}^n (y_i - \bar{y})^2 - b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right] = \frac{n-1}{n} \left(s_y^2 - b^2 s_x^2 \right) \quad (4.39)$$

【例 4.5】(续例 4.2) 利用回归估计量推算该县船舶调查月完成的货运量。

解:根据例 4.2 中的计算结果可得样本回归系数:

$$b = \frac{s_{xy}}{s_x^2} = \frac{23\,382.22}{2\,161.11} = 10.819\,5$$

从而

$$y_{lr} = \bar{y} + b(X - \bar{x}) = 1\,123.2 + 10.819\,5 \times \left(\frac{154\,626}{2\,860} - 65 \right) = 1\,004.89$$

因此,该县船舶调查月完成的货运量的回归估计为:

$$\hat{Y} = N y_{lr} = 2\,860 \times 1\,004.89 = 2\,873\,982(\text{吨})$$

为了估计 \hat{Y}_{lr} 的方差,先计算回归残差方差:

$$s_e^2 = \frac{n-1}{n-2} (s_y^2 - b^2 s_x^2) = \frac{10-1}{10-2} \times (421\,179.07 - 10.819\,5^2 \times 2\,161.11) = 189\,218.52$$

于是 \hat{Y}_{lr} 方差的估计为:

$$v(\hat{Y}_{lr}) = \frac{N^2(1-f)}{n} S_e^2 = 2860^2 \times \left(\frac{1}{10} - \frac{1}{2860} \right) \times 189218.52 \\ = 1.54232 \times 10^{11}$$

\hat{Y}_{lr} 标准差的估计为:

$$s(\hat{Y}_{lr}) = \sqrt{v(\hat{Y}_{lr})} = 392724(\text{吨})$$

与例 4.2 的结果比较,对于本问题,回归估计优于比率估计,而比率估计又优于简单估计。回归估计优于比率估计的原因是回归直线没有通过原点。需要注意的是,为了说明问题,本例样本量不大,在实际工作中,对于样本量较小的情形,必须考虑比率估计及回归估计的偏倚。

对于简单随机抽样,为了比较上述比率估计量、回归估计量及简单估计量的优劣,可通过比较它们的均方误差或方差大小来进行。简单估计量是无偏的,而比率估计量和回归估计量是渐近无偏的,因此这里只比较当 n 比较大的情形时,估计量的方差大小。

由上面的讨论,我们知道:

$$V(y) = \frac{1-f}{n} S_y^2 \\ V(\bar{y}_R) \approx \frac{1-f}{n} (S_y^2 + R^2 S_x^2 - 2RS_{yx}) \\ V(y_{lr}) \approx \frac{1-f}{n} S_y^2 (1 - \rho^2)$$

由此可以看出:

1. 回归估计量总是优于简单估计量,除非 $\rho = 0$, 即

$$V(y_{lr}) \leq V(y) \quad (4.40)$$

2. 比率估计量优于简单估计量的条件是

$$\rho > \frac{1}{2} \frac{\frac{S_x}{\bar{X}}}{\frac{S_y}{\bar{Y}}} - \frac{C_x}{2C_y} \quad (4.41)$$

这时,比率估计量优于简单估计量。

3. 回归估计量优于比率估计量的条件是

$$\rho^2 S_y^2 \leq R^2 S_x^2 - 2R\rho S_y S_x \quad (4.42)$$

也就是

$$(RS_1 - \rho S)^2 \geq 0 \quad (4.43)$$

或者说

$$(B - R)^2 \geq 0 \quad (4.44)$$

因此,除了 $B = R$ 的情况之外,回归估计量总是优于比率估计量。只有当 y_i 与 x_i 的关系式为通过原点的一条直线时,才有 $B = R$ 成立。

四、分层随机抽样下的回归估计

与比率估计类似,分层随机抽样时,如果采用回归估计,则当各层样本量不小时,可先在各层回归估计,然后将各层汇总,得到总体指标的估计,这种方式称为分别回归估计。如果各层样本量不大,则也可采用联合回归估计。

(一) 分别回归估计

对于分层随机抽样,总体均值 \bar{Y} 和总体总量 Y 的分别回归估计量(separate regression estimator)为:

$$y_{lrs} = \sum_{h=1}^L W_h y_{lrh} = \sum_{h=1}^L W_h [y_h + \beta_h (X_h - x_h)] \quad (4.45)$$

$$\hat{Y}_{lrs} = N y_{lrs} = \sum_{h=1}^L N_h [y_h + \beta_h (\bar{X}_h - \bar{x}_h)] \quad (4.46)$$

当各层的回归系数为事先给定的常数时,分别回归估计量是无偏的,其方差为:

$$V(y_{lrs}) = \sum_h \frac{W_h^2 (1 - f_h)}{n_h} (S_{yh}^2 + \beta_h^2 S_{xh}^2 - 2\beta_h S_{yxh}) \quad (4.47)$$

并且当

$$\beta_h = B_h = \frac{S_{yxh}}{S_{xh}^2}, h = 1, 2, \dots, L \quad (4.48)$$

时, $V(y_{lrs})$ 达到最小,即

$$V_{\min}(y_{lrs}) = \sum_{h=1}^L \frac{W_h^2 (1 - f_h)}{n_h} S_{yh}^2 (1 - \rho_h^2) \quad (4.49)$$

通常 β_h 未知,可以用样本回归系数 b_h 作为 β_h 的估计:

$$b_h = \frac{\sum_{i=1}^{n_h} (y_{hi} - y_h)(x_{hi} - x_h)}{\sum_{i=1}^{n_h} (x_{hi} - x_h)^2} \quad (4.50)$$

这时,分别回归估计量是有偏的,但当每一层的样本量 n_h 都较大时,估计的偏倚可

以忽略,其方差近似为:

$$V(y_{lrs}) \approx \sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} S_{eh}^2 (1 - \rho_h^2) \quad (4.51)$$

方差 $V(y_{lrs})$ 的样本估计为:

$$v(y_{lrs}) = \sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} s_{eh}^2 \quad (4.52)$$

式中, $s_{eh}^2 = \frac{1}{n_h - 2} \sum_{i=1}^{n_h} [(y_{hi} - y_h) - b_h(r_{hi} - x_h)]^2$

分别回归估计量要求每一层的样本量都较大,如果这个条件得不到满足,则分别回归估计量的偏倚可能很大,这时,采用联合回归估计量更好些。

(二) 联合回归估计

对于分层随机抽样,总体均值 \bar{Y} 和总体总量 Y 的联合回归估计量(combined regression estimator)为:

$$y_{lrc} = y_{st} + \beta(X - x_{st}) \quad (4.53)$$

$$\hat{Y}_{lrc} = N\hat{y}_{lrc} = \hat{Y}_{st} + \beta(X - \hat{X}_{st}) \quad (4.54)$$

式中, y_{st} 和 x_{st} 分别为 Y 和 X 的分层估计。

对于分层随机抽样联合回归估计量,当回归系数为事先给定的常数时,作为 Y 及 Y 的回归估计, y_{lrc} 及 \hat{Y}_{lrc} 都是无偏的。 y_{lrc} 和 \hat{Y}_{lrc} 的方差为:

$$V(\hat{y}_{lrc}) = \sum_h \frac{N_h^2(1-f_h)}{N^2 n_h} (S_{yh}^2 + \beta^2 S_{xh}^2 - 2\beta S_{yhxh}) \quad (4.55)$$

$$V(\hat{Y}_{lrc}) = \sum_h \frac{N_h^2(1-f_h)}{n_h} (S_{yh}^2 + \beta^2 S_{xh}^2 - 2\beta S_{yhxh}) \quad (4.56)$$

并且,只要 β 取

$$B_c = \frac{\sum_{h=1}^L \frac{W_h^2(1-f_h) S_{yhxh}}{n_h}}{\sum_{h=1}^L \frac{W_h^2(1-f_h) S_{xh}^2}{n_h}} \quad (4.57)$$

时, $V(\hat{y}_{lrc})$ 达到最小。

当回归系数未知时,取 β 为 B_c 的样本估计:

$$b_c = \frac{\sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \sum_{i=1}^{n_h} (y_{hi} - y_h)(x_{hi} - \bar{x}_h)}{\sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2} = \frac{\sum_h \frac{W_h^2(1-f_h)}{n_h} S_{yhxh}}{\sum_h \frac{W_h^2(1-f_h)}{n_h} S_{xh}^2} \quad (4.58)$$

这时联合回归估计是有偏的,但当样本量 n 较大时,估计量的偏倚趋于零,回归估计是渐近无偏的,且

$$V(y_{ln}) \approx \sum_h \frac{W_h^2(1-f_h)}{n_h} (S_{yh}^2 + B_c^2 S_{xh}^2 - 2B_c S_{yhxh}) \quad (4.59)$$

方差 $V(y_{ln})$ 的样本估计为:

$$v(y_{ln}) = \sum_h \frac{W_h^2(1-f_h)}{n_h} (s_{yh}^2 + b_c^2 S_{xh}^2 - 2b_c s_{yhxh}) \quad (4.60)$$

(三) 分别回归估计与联合回归估计的比较

当回归系数事先设定时,分别回归估计优于联合回归估计,尤其在各层回归系数相差较大时,分别回归估计更好。

当回归系数由样本估计时,如果各层的样本量不太小,且各层的回归系数相差较大,还是采用分别回归估计为宜。若各层的样本量不太大,且各层的回归系数大致相同,则采用联合回归估计较好。若层内的回归系数差别不是太大,而每层的样本量并非都相当大时,联合回归估计可能更保险些。

【例 4.6】(续例 4.4) 利用回归估计量估计该市港口生产单位 1997 年完成的吞吐量

解:样本回归系数:

	$h = 1$, 非国有	$h = 2$, 国有
b_h	1.070 17	0.856 402

根据例 4.3 中的计算中间结果,则

(1) 按分别回归估计量估计:

$$\begin{aligned} \hat{Y}_{lrs} &= \sum_{h=1}^2 N_h \bar{y}_{lh} = \sum_{h=1}^2 N_h [y_h + b_h (X_h - \bar{x}_h)] \\ &= 163\,421.10 + 107\,135.19 = 270\,556.30 \\ v(\hat{Y}_{lrs}) &= \sum_{h=1}^2 \frac{N_h^2(1-f_h)}{n_h} \frac{n_h-1}{n_h-2} (s_{yh}^2 + b_h^2 s_{xh}^2) \\ &= 70\,809\,522.4 + 19\,062\,946.81 = 89\,872\,469.22 \end{aligned}$$

$$s(\hat{Y}_{lr}) = \sqrt{v(\hat{Y}_{lr})} = 9\,480.11$$

(2) 按联合回归估计量估计:

$$b = \frac{\sum_{h=1}^2 \frac{W_h^2(1-f_h)}{n_h} s_{ywh}}{\sum_{h=1}^2 \frac{W_h^2(1-f_h)}{n_h} s_{xwh}} = \frac{756.575\,7}{735.253\,5} = 1.029\,0$$

$$\hat{Y}_r = \hat{Y}_d + b_c(X - \hat{X}_d)$$

$$277\,310 + 1.029 \times (274\,300 - 279\,700) = 271\,753.4$$

$$\begin{aligned} v(\hat{Y}_{lr}) &= \sum_{h=1}^2 \frac{N_h^2(1-f_h)}{n_h} (s_{dh}^2 + b_c^2 s_{wh}^2 - 2b_c s_{ywh}) \\ &= 63\,849\,916.5 + 21\,508\,415.67 = 85\,358\,332.17 \end{aligned}$$

$$s(\hat{Y}_{lr}) = \sqrt{v(\hat{Y}_{lr})} = 9\,238.96$$

§ 4.4 差值估计

如果调查时所用的辅助变量为目标量最近的普查结果,或者回归估计的回归系数接近于1,这时可以采用差值估计。

对于简单随机抽样,总体均值的差值估计量(difference estimator)为:

$$\begin{aligned} y_d &= \bar{y} + X - \bar{x} \\ &= X + (y - x) = X + d \end{aligned} \quad (4.61)$$

$$\text{式中, } d = y - \bar{x} = \frac{1}{n} \sum_{i=1}^n (y_i - x_i)$$

性质3 作为 $\beta = 1$ 的回归估计量, y_d 是 Y 的无偏估计,其方差为:

$$V(y_d) = \frac{1-f}{n} (S_y^2 + S_x^2 - 2S_{yx}) \quad (4.62)$$

将 S_y^2, S_x^2, S_{yx} 的样本估计值代入上式,就可得到 $V(y_d)$ 的样本估计。

【例4.7】(续例4.4) 利用差值估计量估计该市港口生产单位1997年完成的吞吐量,并比较估计量及其精度。

解:由于辅助变量是上年的普查资料,且样本回归系数接近于1,因此可以采用差值估计。

$$\hat{Y}_d = \sum_{h=1}^2 N_h [y_h + (X_h - x_h)] = \hat{Y}_{st} + X - \hat{X}_q$$

$$\begin{aligned}
&= 277\,310 + 274\,300 - 279\,700 - 271\,910 \\
v(\hat{Y}_d) &= \sum_{h=1}^L N_h^2 (1 - f_h) \left(s_{yh}^2 + s_{xh}^2 - 2s_{xyh} \right) \\
&= 65\,579\,831.11 + 20\,336\,554.29 - 85\,916\,385.4 \\
s(\hat{Y}_d) &= \sqrt{v(\hat{Y}_d)} = 9\,269.11
\end{aligned}$$

最后,为比较本例所用的五种估计量,将估计量、估计量标准差的估计、估计量的变异系数列于表 4.7。

表 4.7

估计方法	总量估计	标准差估计	估计量的变异系数
分别比率估计量	272 536.6	9 588.48	0.035 182
联合比率估计量	271 956.1	9 289.44	0.034 158
分别回归估计量	270 556.3	9 480.11	0.035 039
联合回归估计量	271 753.4	9 238.96	0.033 998
差值估计	271 910.0	9 269.11	0.034 089

对于本例,从数值上看,五种估计量的精度非常接近,回归估计量比相应的比率估计量的标准差要小,联合回归估计量的标准差最小,但通常用样本回归系数的回归估计量的偏倚比比率估计量的大,且本问题样本量较小,因此从 MSE 的角度,回归估计量未必是最好的,由于差值估计量是无偏的,且其标准差与比率估计量、回归估计量相当,因此对本问题,差值估计量是最优的。

对于简单随机抽样,简单估计、差值估计是无偏的,比率估计、回归估计是渐近无偏的。当样本量较小时,不能忽略比率估计及回归估计的偏倚,而在小样本时,回归估计的偏倚可能比比率估计的大,因此,从均方误差的意义来看,这时回归估计并不一定比比率估计好。

当辅助变量为调查指标的最近的普查值时,可以考虑使用差值估计,尽管差值估计的方差可能比回归估计要大,但由于它是无偏估计,所以,总的均方误差可能比回归估计的小。

小 结

本章介绍了简单随机抽样比率估计量、回归估计量及其性质。比率估计量除了

用于对总体比率量进行估计外,在实际工作中,人们使用比率估计量和回归估计量主要是利用辅助变量提高估计效率。

比率估计量和回归估计量是有偏的,但当样本量足够大时,其偏倚可以忽略。与简单随机抽样简单估计量相比,只要辅助变量与调查指标相关性较好,就能保证比率估计量、回归估计量比简单估计量有效。比率估计量、回归估计量不仅可以用于简单随机抽样,也可以用于分层随机抽样。

使用比率估计量、回归估计量以提高估计精度时要求已知辅助变量的总体总量或总体均值,如果辅助变量的总体总量或总体均值未知,则要采用二重抽样,以解决辅助信息不足的问题。

本章附录 比率估计量、回归估计量性质的证明

1. 证明比率估计的偏倚。

证明:(1) 比率估计的近似偏倚。

$$\hat{R} = R + \frac{y - Rx}{x}$$

其中

$$\frac{1}{x} = \frac{1}{\bar{X} + (x - \bar{X})} = \frac{1}{\bar{X}} \left(1 + \frac{x - \bar{X}}{\bar{X}} \right)^{-1}$$

对其用泰勒级数公式展开,得到

$$\begin{aligned} \frac{1}{\bar{X}} \left(1 + \frac{x - \bar{X}}{\bar{X}} \right)^{-1} &= \frac{1}{\bar{X}} \left[1 - \frac{x - \bar{X}}{\bar{X}} + \left(\frac{x - \bar{X}}{\bar{X}} \right)^2 - \dots \right] \\ &\approx \frac{1}{\bar{X}} \left(1 - \frac{x - \bar{X}}{\bar{X}} \right) \end{aligned}$$

因此

$$\hat{R} - R \approx \frac{y - Rx}{\bar{X}} \left(1 - \frac{x - \bar{X}}{\bar{X}} \right) = \frac{\bar{y} - R\bar{x}}{\bar{X}} - \frac{(y - Rx)(x - \bar{X})}{\bar{X}^2}$$

由于 $E(y - Rx) = Y - RX = 0$

因而偏倚的主要项来自于等式右边的第二项。由

$$\begin{aligned} E[\bar{y}(\bar{x} - \bar{X})] &= E[(y - \bar{Y})(\bar{x} - \bar{X})] = \frac{1}{n} f S_{yx} = \frac{1}{n} \rho S_y S_x \\ E[x(x - \bar{X})] &= E(x - \bar{X})^2 = \frac{1-f}{n} S_x^2 \end{aligned}$$

因此, 偏倚的主要项为

$$E(\hat{R} - R) \approx \frac{1}{X^2} [E_y(x - X) + RE_x(r - X)] \\ + \frac{1-f}{nX^2} (RS_d^2 - \rho S_y S_x)$$

(2) 比率估计的精确偏倚。

考虑 \hat{R} 和 x 的协方差:

$$\text{Cov}(x, \hat{R}) = E\left(\frac{y}{x}\right) - E(\hat{R})E(x) = Y - XE(\hat{R})$$

因此

$$E(\hat{R}) = \frac{Y}{X} - \frac{1}{X} \text{Cov}(x, \hat{R}) = R - \frac{1}{X} \text{Cov}(x, \hat{R})$$

从而

$$E(\hat{R}) = R - \frac{1}{X} \text{Cov}(x, \hat{R})$$

2. 证明比率估计的近似方差。

证明: 因为 \hat{Y}_R, y_R 与 \hat{R} 只差一个常数, 这里只给出对 \hat{R} 近似方差的证明。

$$\hat{R} - R = \frac{y}{x} - R = \frac{y - Rx}{x}$$

当 n 足够大时, $x \approx X$, 将其代入上式分母, 得

$$\hat{R} - R \approx \frac{y - Rx}{X}$$

于是

$$E(\hat{R} - R) \approx \frac{1}{X} [E(y) - R \cdot E(x)] = \frac{1}{X} (Y - R \cdot X) = 0$$

因此, 当 n 足够大时, $E(\hat{R}) \approx R$ 。这时

$$V(\hat{R}) \approx \text{MSE}(\hat{R}) = E(\hat{R} - R)^2 \approx \frac{1}{X^2} E(y - Rx)^2$$

注意到 $y - Rx$ 是 $d_i = y_i - Rx_i$ 的样本均值, 且 d_i 的总体均值 $D = Y - RX = 0$, 因此

$$V(\hat{R}) \approx \frac{1}{X^2} E(d)^2 = \frac{1}{X^2} \cdot \frac{1-f}{n} S_d^2 \\ = \frac{1}{X^2} \cdot \frac{1-f}{n} \cdot \frac{1}{N-1} \sum_{i=1}^N D_i^2 = \frac{1}{X^2} \cdot \frac{1-f}{n} \cdot \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2$$

为估计 $V(\hat{R})$, 用

$$\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2$$

估计 $\frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2$ 。这个估计也是有偏的, 但当 n 足够大时, 估计的偏倚趋于零。因此 $V(\hat{R})$ 的估计为:

$$\begin{aligned} v_1(\hat{R}) &= \frac{1}{X^2} \frac{1-f}{n} \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 \\ &= \frac{1-f}{nX^2} \frac{1}{n-1} \sum_{i=1}^n (y_i^2 - 2\hat{R}y_ix_i + \hat{R}^2x_i^2) \\ &= \frac{1-f}{nX^2} (s_y^2 - 2\hat{R}s_{xy} + \hat{R}^2s_x^2) \end{aligned}$$

对 $V(\hat{R})$ 的估计式中, 也可以用 \bar{x} 代替 X , 得到 $V(\hat{R})$ 另一种估计式:

$$v_2(\hat{R}) = \frac{1-f}{n\bar{x}^2} (s_y^2 - 2\hat{R}s_{xy} + \hat{R}^2s_x^2)$$

3. 证明 β 为常数时回归估计的性质。

证明: 记 $\beta = \beta_0$, 下面给出对总体均值的回归估计量的性质的证明。这时

$$y_{lr} = y - \beta_0(x - X)$$

因此

$$E(y_{lr}) = E(y) - \beta_0[E(\bar{x}) - \bar{X}] = \bar{Y}$$

为求 y_{lr} 的方差, 可以将 y_{lr} 看做 $y_i - \beta_0(x_i - \bar{X})$ 的样本均值, 因此由简单随机抽样简单估计量的方差公式, 可以得到

$$\begin{aligned} V(y_{lr}) &= \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^N [Y_i - \beta_0(X_i - \bar{X}) - \bar{Y}]^2 \\ &= \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - \beta_0(X_i - \bar{X})]^2 \\ &= \frac{1-f}{n} (S_y^2 + \beta_0^2 S_x^2 - 2\beta_0 S_{xy}) \end{aligned}$$

将 β_0 看做变量 Z , 由 $aZ^2 + bZ + c$ 当 $Z = -\frac{b}{2a}$ 且 $a > 0$ 时达到极小值, 这里 $a = S_x^2 > 0$, 因此当

$$\beta_0 = -\frac{2S_{xy}}{2S_x^2} = -\frac{S_{xy}}{S_x^2}$$

时, $V(y_{lr})$ 达到极小值:

$$V_{\min}(y_{lr}) = \frac{1}{n} \frac{f}{f} \left(S_y^2 - \frac{S_{yx}^2}{S_x^2} \right) = \frac{1}{n} f S_y^2 (1 - \rho^2)$$

由于 s_y^2, s_x^2, s_{yx} 是 S_y^2, S_x^2, S_{yx} 的无偏估计, 将它们代入上式, 即可得 $V(y_{lr})$ 的无偏估计:

$$v(y_{lr}) = \frac{1-f}{n} (s_y^2 + \beta_0^2 s_x^2 - 2\beta_0 s_{yx})$$

4. 证明大样本时, 比较三种估计量优劣的条件。

证明: 大样本时, 有

$$V(y) = \frac{1-f}{n} S_y^2$$

$$V(y_R) \approx \frac{1}{n} f (S_y^2 + R^2 S_x^2 - 2RS_{yx})$$

$$V(y_{lr}) \approx \frac{1-f}{n} S_y^2 (1 - \rho^2)$$

由于 $\rho^2 \leq 1$, 因此, 除 $\rho = 0$ 以外, 总有

$$V(y_{lr}) \leq V(y)$$

比率估计量优于简单估计量的条件是

$$S_y^2 > S_y^2 + R^2 S_x^2 - 2RS_{yx}$$

即 $2RS_{yx} > R^2 S_x^2$

$$\frac{S_{yx}}{S_x^2} > \frac{R}{2}$$

于是有

$$\rho > \frac{1}{2} \frac{\frac{S_x}{\bar{X}}}{\frac{S_y}{\bar{Y}}} = \frac{C_x}{2C_y}$$

回归估计量优于比率估计量的条件是

$$S_y^2 (1 - \rho^2) \leq S_y^2 + R^2 S_x^2 - 2RS_{yx}$$

这等价于

$$(\rho S_y - RS_x)^2 \geq 0$$

即 $(B - R)^2 \geq 0$

因此, 除 $B = R$ 的情况之外, 回归估计量总是优于比率估计量。

习 题

1. 从一个总体中抽出一个简单随机样本,对样本中每个单元测量了 y 和 x 的值。若 x 的总体均值 X 已知,在下面的方法中,你选择那一种方法估计 $\frac{Y}{X}$? 并说明你

的理由 (1) 总是用 $\frac{y}{X}$; (2) 有时用 $\frac{y}{X}$, 有时用 $\frac{y}{x}$; (3) 总是用 $\frac{y}{x}$ 。

2. 对某十字路口车流量进行观测,试判断如下一些量的类型。

- (1) 一周内通过路口的车辆数;
- (2) 一周内通过路口的小轿车的比例;
- (3) 路口每秒通过的车辆数;
- (4) 车辆在路口的平均等待时间;
- (5) 本地车牌照尾数为奇数和偶数的车辆通过数之比。

并请你举例说明你知道的总体各种类型的量。

3. 找一本英汉词典,从前言找到该词典收录的词条数,从目录可以查出正文的页数,以正文的每一页作为抽样单元,利用随机数表,在正文中随机抽取 $n = 50$ 页,记录被抽中页中你认识的单词数以及该页包括的单词数,试用适当的估计方法估计你的英语词汇量,给出估计的精度,并说明你选择估计方法的理由。如果要求在 95% 置信度下,估计的相对误差不超过 10%,则应该抽取多少页?

4. 某市欲估计居民用于购买书报杂志的支出占总收入的比重,在全体居民户中随机抽出 20 户居民,调查了样本居民户最近一年的购买书报杂志的支出 y_i (元) 及家庭总收入 x_i (百元),结果如下:

i	y_i	x_i	i	y_i	x_i
1	550	300	11	150	242
2	370	291	12	350	265
3	200	289	13	230	254
4	120	223	14	250	245
5	160	201	15	480	305
6	320	317	16	390	303
7	290	279	17	210	267
8	70	180	18	380	277
9	90	189	19	230	227
10	110	205	20	420	271

试估计该市居民家庭每百元收入用于购买书报杂志的支出,并计算估计量的标准差

5. 某公司欲了解广告对其产品销售量的作用,从销售该公司产品的 452 家企业中抽选了 20 家,分别调查了广告前与广告后的月销售量数据,如下表。

样本企业	广告前	广告后	样本企业	广告前	广告后
1	208	239	11	599	626
2	400	428	12	510	538
3	440	472	13	828	888
4	259	276	14	473	510
5	351	363	15	924	998
6	880	942	16	110	171
7	273	294	17	829	889
8	487	514	18	257	265
9	183	195	19	388	419
10	863	897	20	244	257

(1) 若广告前的月总销售量为 216 256, 分别用比率估计量和回归估计量估计广告后的月总销售量及其标准差;

(2) 求广告后比广告前销售量增加百分比 95% 的置信区间;

(3) 若允许估计总销售量的绝对误差为 $\Delta = 3\ 800$, 置信度为 95%, 确定应抽多少样本企业

6. 对于如下表样本, 如果 X 已知, 准备采用比率估计量对 Y 进行估计, 你选择哪个 X 作为辅助变量?

i	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	y	i	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	y
1	57	17	211	113	11	50	12	171	172
2	5	1	0	1	12	8	3	32	8
3	10	2	0	1	13	22	12	497	36
4	9	6	120	30	14	20	3	30	15
5	15	3	0	0	15	7	5	0	18
6	15	5	30	26	16	23	5	40	17
7	5	2	5	6	17	14	4	70	25
8	22	4	330	10	18	12	6	40	10
9	18	5	272	30	19	7	2	15	1
10	26	10	70	0	20	6	2	1	2

7. 下列数据是 $N = 6$, 两层单元数相同的人为总体。

第 一 层		第 二 层	
x_{1i}	y_{1i}	x_{2i}	y_{2i}
3	0	8	6
5	3	15	9
10	6	25	15

对 $n_1 = n_2 = 2$ 的一个分层随机样本：

- (1) 列出所有可能的样本；
- (2) 用分别比率估计和联合比率估计估计 Y , 计算估计的偏差及 MSE ；
- (3) 对上述结果进行比较分析。



第 5 章

不等概抽样

前面学习的是等概率抽样方法,即每个单元入样的概率相等。与此相对应的另一类方法是不等概抽样,也即每个单元入样的概率不相等。

本章共分三节,第一节介绍不等概抽样的概念,第二节介绍与单元大小成比例的放回的不等概抽样(PPS 抽样),第三节介绍与单元大小成比例的不放回不等概抽样(π PS 抽样)。

§ 5.1 引言

一、概念与特点

前面所学的简单随机抽样,总体中的每个单元具有同样的入样概率,它是等概率抽样,在分层随机抽样时,层内单元是按简单随机抽样抽取的,因此,层内也是等概率抽样。等概率抽样的特点是总体中每个单元地位相同,在抽样时对每个单元采取“不偏不倚”的态度。

与等概率抽样对应的另一类方法是不等概抽样,也就是在抽样前赋予总体每个单元一个入样概率,当然这个入样概率是不相同的,否则抽样就成为等概率的

抽样。

当总体单元之间差异不大时,简单随机抽样是简便的、有效的。例如,对家庭消费支出的调查中,以家庭为抽样单元,由于家庭之间的差异不是很大,因此用简单随机抽样是有效的。

当总体单元之间差异非常大时,简单随机抽样效果并不好。例如,对船舶运输量进行调查时,以船舶为抽样单元,则有的是从事远洋运输的万吨巨轮,更多的是从事内河河网地区运输的上百吨乃至几十吨小船,这时,简单随机抽样的效果肯定不好。又如,对某市商业销售额进行调查时,以商场为抽样单元,则少数商场是大型或特大型的,而多数是中小商店,这时,简单随机抽样效果也不好。对于这两种情况,人们自然会想到,将大船(大型商场)和小船(小商店)同等对待并不合理,因为大船(大型商场)数量虽然不多,但占总运输量(销售额)的份额较大;另外,由于规模和管理水平的原因,往往大船(大型商场)的调查比较容易,可以做得细致一些,而小船(小商店)的调查往往比较困难,也没有必要对占市场份额不大的这部分单元花太大的精力做过多的调查。因此在调查时,大船(大型商场)应该处于更重要的地位。

出现总体单元差异特别大的情况时,通常是牺牲“简单”来提高抽样效率。一种做法是将总体单元按规模(大小)分层,对较大单元的层抽样比定得高些,抽样比甚至可以是100%,而较小单元的层抽样比定得低一些。另一种做法就是赋予每个单元与其规模(或辅助变量)成比例的人样概率,这样一来,大单元入样概率大,小单元入样概率小。

不等概抽样时,总体中某类单元比其他单元出现在样本中的机会大,这给人一种感觉,这部分单元对推算的影响大,使得推算偏向于某一方。例如,大商场抽得多了,会不会使得推算的销售额偏大。但事实上,某些单元入样概率较大,推算时,则赋予其一个较小的权;反之,入样概率较小,推算时,则赋予其一个较大的权,从而使推算结果仍然是公平的。

实际工作中,如果遇到下面几种情况,则可以考虑使用不等概抽样。

1. 抽样单元在总体中所占的地位不一致。例如上面讨论的船舶、商场等调查问题。

2. 调查的总体单元与抽样总体的单元不一致。例如某大型单位准备对职工家庭情况进行调查,一种自然的办法是以人事部门的职工花名册作为抽样框进行抽样。该单位有少数家庭两名职工在该单位工作,如果对职工进行简单随机抽样,则双职工家庭被抽中的概率大,而调查者希望对家庭进行等概率抽样。除了对抽样框进行整理,将双职工家庭中的一名成员从抽样框中拿掉以外,可以对职工采用不等

概抽样,一种做法是对每名职工记录其家庭成员在该单位工作的人数,然后对每名职工按与人数成反比的概率进行抽样

3. 改善估计量 不等概抽样可用于对估计量进行改善,例如简单随机抽样比率估计量是渐进无偏的,要使它成为无偏估计,只要每个大小为 n 的样本被抽中的概率与其辅助变量的和 $\sum_{i=1}^n x_i$ 成比例(如水野法),则这时的比率估计量就是无偏估计量,而这个样本并不是简单随机样本,而是一个不等概抽样获得的样本。

不等概抽样除了应用于上述几种情况外,还广泛应用于整群抽样、多阶段抽样中群或初级单元大小相差较大的情形

不等概抽样的优点主要是大大提高估计精度,减少抽样误差,但使用它也有条件,就是必须要有说明每个单元规模大小的辅助变量来确定每个单元入样的概率,这在抽样及推算时都是必须的。有时,对应于每个单元的辅助变量的获得比较容易或方便,例如,管理部门在车船登记台账中,车船名及其载重吨位是同时登记的,以载重吨位作为辅助变量时,抽样框的编制几乎与简单随机抽样一样方便。但对有些问题要复杂一些,例如将某县的农田划分成地块后,以地块的面积作为辅助变量,则这时除了对地块进行编号,还要对地块的面积进行丈量。因此,同简单随机抽样相比,不等概抽样编制抽样框的过程有时要复杂一些。

二、不等概抽样的种类

布鲁尔(Brewer)和哈尼夫(Hanif)在《不等概率抽样》(1983)中列举了50多种不等概抽样方法,但常用的大约10种。对不等概抽样的分类可以有多种原则,可以按样本单元是否放回分为放回不等概抽样和不放回不等概抽样。

(一) 放回不等概抽样

每次在总体中对每个单元按入样概率进行抽样,抽取出来的样本单元放回总体,然后进行下一次抽样,这样的话,每次抽样过程都是从同一个总体独立进行的。放回不等概抽样实施及推算过程相对来说比不放回的简单。

(二) 不放回不等概抽样

每次在总体中对每个单元按入样概率进行抽样,抽取出来的样本单元不再放回总体,对总体中剩下的单元进行下一次抽样。不放回不等概抽样的效率比放回时的效率高,但是不放回不等概抽样的实施及推算过程比放回时复杂得多。

对于不放回不等概抽样,样本的抽取可以有以下几种方法。

1. 逐个抽取法。每次从总体未被抽中的单元中以一定的概率抽取一个样本单元,通常这个概率与已被抽中的样本单元有关。

2. 重抽法: 以一定的概率逐个进行放回抽样, 如果抽到重复单元, 则放弃所有抽到的样本单元, 重新抽取, 直至抽到规定的样本量且所有样本单元不重复。

3. 全样本抽取法。对总体每个单元分别按一定概率决定其是否入样。这种方法的样本量是随机的, 事先不能确定, 而且它可能出现总体中全体单元都入样或全都未入样。

4. 系统抽样法。将总体单元按某种顺序排列, 将规定的人样概率汇总, 根据样本量确定抽样间距 k , 在 $1 \sim k$ 产生一个随机数, 并确定相应的初始单元, 以后在总体中每隔 k 个单元抽出一个作为样本单元。

三、区域抽样

区域抽样(area sampling) 也称为面积抽样。这种方法主要用于以下的情形: 区域或面积本身就是抽样单元, 或者抽样单元的名单抽样框无法获得, 但每个抽样单元只隶属于某个区域。例如, 某县进行小麦产量调查时, 将全县农田土地按易于划分的规则划分成地块(如利用沟垄、水渠、道路等地理特征自然隔离), 然后对地块进行抽样, 对被抽中地块的小麦产量进行实割实测, 从而推算全县的产量。由于地块的面积通常不相等, 因此对地块的抽样可以是简单随机抽样, 也可以按地块的面积进行不等概抽样。

为此, 需要对抽样框类型进行讨论。抽样框可以分为名单抽样框和区域抽样框。

名单抽样框由抽样单元的名单组成。例如, 某高校全体在校学生的花名册就是一个名单抽样框。又如, 在工商管理部门登记的企业名册也是一个名单抽样框。

区域抽样框由定义明确的区域组成, 而一个区域是由个体组成的。例如, 我们对居民家庭进行某项调查时, 可以利用地图编制各行政区的名单, 或到街道办事处获得居委会的名单, 这时的行政区及居委会都是由个体(居民户)组成的区域。又如将农田土地划分成地块。

对于区域我们可以直接进行抽样, 这时的抽样单元就是区域本身, 例如对地块的抽样。

大多数情况下, 抽样单元是区域内的个体, 这时有两种选择, 即对区域内的所有单元进行调查, 或者对区域内的单元再抽样, 它们分别是后面将要介绍的整群抽样和多阶段抽样。

一般来说, 抽样调查的总体比较大, 要编制全体抽样单元的名单往往很困难, 而且也没有必要。这时比较容易的做法是通过对区域的划分, 建立区域抽样框, 然后对被抽中的区域进行调查, 或者再编制下一阶段的抽样框。如果有必要, 这个抽

样框也可以是 π 域抽样框

例如,对北京市中学生的某项调查,没有必要将全体在校生的名单都拿来,可以对学校进行抽样。对被抽中的学校,可以直接利用学生处的学生名单进行抽样,但对于较大的学校可能还是不方便,因此可以抽学生班级并对被抽中班级的全体学生进行调查或对班级中的学生再抽样。

区域抽样框有以下主要优点:

(1) 容易定义和识别 区域抽样框很容易通过地图或行政区划加以定义,而且能很清楚地识别

(2) 比较稳定 区域相对来说比较稳定。例如,我们调查一个居民楼中的所有居民户,比利用居民户名单抽样框要容易得多,因为前者是稳定的,而后者可能在调查的时候已经搬迁

(3) 容易操作,回答率较高。现场工作人员能很容易并清楚地识别和确定区域的界限,从而比较容易地找到样本单元,使回答率提高。

§ 5.2 放回不等概抽样

一、PPS 抽样

(一) 多项抽样与 PPS 抽样

设 Z_1, Z_2, \dots, Z_N 是一组概率, $\sum_{i=1}^N Z_i = 1$, 按这组概率对总体中的 N 个单元进行放回抽样, 每次抽中第 i 个单元的概率为 Z_i , 独立地进行这样的抽样 n 次, 则这种不等概抽样为多项抽样。

特别地, 如果每个单元有说明其大小或规模的度量 M_i , 则 Z_i 可取

$$Z_i = \frac{M_i}{M_0} = \frac{M_i}{\sum_{i=1}^N M_i} \quad (5.1)$$

这时, 每个单元在每次抽选中入样的概率与其单元规模的大小成比例, 因而多项抽样称为放回的与单元规模大小成比例的概率抽样 (sampling with probability proportional to size), 简称 PPS 抽样。

由于抽样是放回的, 因此, 某个单元可能在样本中出现多次, 出现这种情况时, 对这个单元的调查只进行一次, 但计算时按抽中几次计算几次的原则进行。

(二) 实施方法

不等概抽样的实施有两种方法: 代码法与拉希里 (Lahiri) 法。

1. 代码法。在 PPS 抽样中,赋予每个单元与 M_i 相等的代码数,将代码数累加得到 M_0 ,每次抽样都产生一个 $[1, M_0]$ 之间的随机数,设为 m ,则代码 m 所对应的单元被抽中。

如果 M_i 不是整数,则乘以某个倍数。对于一般的多项抽样,通常可以找到某个 M_0 ,使 $M_0 Z_i$ 为整数,每个单元赋予与 $M_0 Z_i$ 相等的代码数,然后进行抽样

【例 5.1】 设某个总体有 $N = 10$ 个单元,相应的单元大小 M_i 及其代码数如表 5.1,我们要在其中产生一个 $n = 3$ 的样本。

表 5.1 利用代码进行 PPS 抽样

i	M_i	$M_i \times 10$	累计 $M_i \times 10$	代码
1	0.6	6	6	1 ~ 6
2	14.5	145	151	7 ~ 151
3	1.5	15	166	152 ~ 166
4	13.7	137	303	167 ~ 303
5	7.8	78	381	304 ~ 381
6	15	150	531	382 ~ 531
7	10	100	631	532 ~ 631
8	3.6	36	667	632 ~ 667
9	6	60	727	668 ~ 727
10	1.1	11	738	728 ~ 738
Σ	$M_0 = 73.8$	738	-	-

先在 $[1, 738]$ 中产生第一个随机数为 354, 再在 $[1, 738]$ 中产生第二个随机数为 553, 最后在 $[1, 738]$ 中产生第三个随机数为 493, 则它们所对应的第 5, 7, 6 号单元被抽中。

2. 拉希里法 令 $M^* = \max_{1 \leq i \leq N} m_i$, 即所有 M_i 中最大值, 每次抽样都分别产生一个 $[1, N]$ 之间的随机数 i 及 $[1, M^*]$ 之间的随机数 m , 如果 $M_i \geq m$, 则第 i 个单元被抽中; 否则, 重抽一组 (i, m) 。

在例 5.1 中, $M^* = 150, N = 10$ 。在 $[1, 10]$ 和 $[1, 150]$ 中分别产生 (i, m) :

$(3, 121), M_3 = 15 < m = 121$, 舍弃, 重抽;

$(8, 50), M_8 = 36 < m = 50$, 舍弃, 重抽;

$(7, 77), M_7 = 100 \geq m = 77$, 第 7 号单元入样;

(5,127), $M_5 = 78 < m = 127$, 舍弃, 重抽;

(4,77), $M_4 = 137 \geq m = 77$, 第4号单元入样;

(9,60), $M_9 = 60 \geq m = 60$, 第9号单元入样。

因此, 第4, 7, 9号单元被抽中。

当样本量 N 很大时, 采用拉希里法不用列出如表 5.1 那样的表, 在这点上, 此法有便捷之处。

二、汉森 赫维茨估计量

对于放回不等概抽样, 对总体总量 Y 的估计是汉森 - 赫维茨 (Hansen - Hurwitz) 估计:

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i} = \frac{M_0}{n} \sum_{i=1}^n \frac{y_i}{m_i} \quad (5.2)$$

\hat{Y}_{HH} 的方差为:

$$V(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^n Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2 \quad (5.3)$$

$V(\hat{Y}_{HH})$ 的无偏估计为:

$$v(\hat{Y}_{HH}) = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{Y}_{HH} \right)^2 = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{m_i} - \frac{\hat{Y}_{HH}}{M_0} \right)^2 \quad (5.4)$$

【例 5.2】 某部门要了解所属 8 500 家生产企业当月完成的利润, 该部门手头已有一份上年各企业完成产量的报告, 将其汇总得到所属企业上年完成的产量为 3 676 万吨。考虑到时间紧, 准备采用抽样调查来推算当月完成的利润。根据经验, 企业的产量和利润相关性比较强, 且企业的特点是规模和管理水平差异比较大, 通常大企业的管理水平较高, 因此采用与上年产量成比例的 PPS 抽样, 从所属企业中抽出一个样本量为 30 的样本, 调查结果如表 5.2。

表 5.2 样本单元的有关数据

i	m_i	y_i	i	m_i	y_i	i	m_i	y_i
1*	38.23	10 926	10	6.50	1 900	19	1.50	10
2	13.70	1 024	11	15.00	864	20	8.00	80
3	0.75	13	12	7.00	17	21	28.42	13 672
4	2.85	30	13	16.00	1 045	22*	9.01	3 845

续前表

i	m_i	y_i	i	m_i	y_i	i	m_i	y_i
5	2.00	1 102	14	12.30	220	23	0.75	480
6	5.00	600	15	3.86	4 600	24	6.00	311
7	10.80	290	16	15.80	2 370	25	28.43	9 284
8	2.00	430	17	9.00	940	26	9.97	842
9	8.81	992	18*	21.00	640	27	6.20	510

* 该样本单位被抽中两次, m_i 为企业当年完成的产量(单位, 万吨); y_i 为企业当月完成的利润(单位, 百元)

要根据以上调查结果估计该部门所属企业当月完成的利润, 并给出 95% 置信度下估计的相对误差。如果要求在相同条件下相对误差达到 20%, 所需的样本量应该是多少?

解: 由上述条件知

$$n = 30, M_0 = 3\,676$$

估计当月完成的利润

$$\begin{aligned}\hat{Y}_{HH} &= \frac{M_0}{n} \sum_{i=1}^n \frac{y_i}{m_i} \\ &= \frac{3\,676}{30} \left(\frac{10\,926}{38.23} + \frac{10\,926}{38.23} + \frac{1\,024}{13.70} + \cdots + \frac{510}{6.2} \right) \approx 757\,087 (\text{百元})\end{aligned}$$

\hat{Y}_{HH} 方差及标准差的估计:

$$\begin{aligned}v(\hat{Y}_{HH}) &= \frac{M_0^2}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{m_i} - \frac{\hat{Y}_{HH}}{M_0} \right)^2 \\ &= \frac{3\,676^2}{30 \times 29} \left[\left(\frac{10\,926}{38.23} - \frac{757\,087}{3\,676} \right)^2 \times 2 + \left(\frac{1\,024}{13.70} - \frac{757\,087}{3\,676} \right)^2 \right. \\ &\quad \left. + \cdots + \left(\frac{510}{6.2} - \frac{757\,087}{3\,676} \right)^2 \right] \\ &\approx \frac{3\,676^2}{30} \times 67\,306.428\,6 \\ &\approx 30\,317\,005\,145.8\end{aligned}$$

$$s(\hat{Y}_{HH}) = \sqrt{v(\hat{Y}_{HH})} \approx 174\,118 (\text{百元})$$

在置信度为 95% 时, 对应的 $t = 1.96$, \hat{Y}_{HH} 的相对误差

$$r = t \frac{s(\hat{Y}_{HH})}{\hat{Y}_{HH}} = 1.96 \times \frac{174\,118}{757\,087} \approx 45\%$$

因此,在置信度仍为 95%、相对误差 $r_1 = 20\%$ 时,所需样本量为:

$$n_1 = \frac{r^2}{r_1^2} n = \left(\frac{0.45}{0.2} \right)^2 \times 30 = 152$$

§ 5.3 不放回不等概抽样

一、 π PS 抽样

(一) 不放回不等概抽样

对于放回抽样,对总体参数的估计及其方差估计比较简单,但样本单元中可能有单元被抽中多次,直观上看,没有必要对同一个单元调查多次,因此放回抽样得到的样本代表性比不放回抽样差。类似于对简单随机抽样的讨论,在同样样本量的条件下,放回抽样的估计精度较低,尤其当抽样 f 比不能忽略时。称不放回的与单元大小成比例的概率抽样为 π PS 抽样

(二) 包含概率

在不放回不等概抽样中,每个单元入样的概率 π_i 及任意两个单元同时入样的概率 π_{ij} 统称为包含概率

对固定的 n , 包含概率满足:

$$\sum_{i=1}^N \pi_i = n \quad (5.5)$$

$$\sum_{j \neq i}^N \pi_{ij} = (n-1)\pi_i \quad (5.6)$$

$$\sum_{i=1}^N \sum_{j>i}^N \pi_{ij} = \frac{1}{2} n(n-1) \quad (5.7)$$

特别地,如果每个单元入样概率与其大小或规模的度量 M_i 严格成比例,记

$Z_i = \frac{M_i}{M_0}$, 则对于固定的 n , 有

$$\pi_i = nZ_i \quad (5.8)$$

这时,我们简称这种情形的抽样为严格的 π PS 抽样。

严格的 π PS 抽样实施起来非常复杂, π_{ij} 不易求得,因此方差的估计也相当困难。严格的 π PS 抽样只有在 $n=2$ 时才有一些简单实用的方法,对于 $n>2$ 的情形,严格的 π PS 抽样则相当复杂。在实际工作中,可以通过分层,在每层中进行严格的 $n=2$ 的 π PS 抽样。

二、霍维茨 汤普森估计量

对于不放回不等概抽样,对总体总量 Y 的估计是霍维茨 汤普森(Horvitz Thompson)估计:

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} \quad (5.9)$$

式中, π_i 为第 i 个单元的包含概率

如果 $\pi_i > 0 (i = 1, 2, \dots, N)$, 则 \hat{Y}_{HT} 是 Y 的无偏估计, 它的方差为:

$$V(\hat{Y}_{HT}) = \sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} Y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{\pi_i - \pi_j}{\pi_i \pi_j} Y_i Y_j \quad (5.10)$$

进一步, 如果 n 固定, 则

$$V(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j>i}^n (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 \quad (5.11)$$

由方差表达式可知, 要使估计量的方差 $V(\hat{Y}_{HT})$ 小, 应尽可能使 $\frac{Y_i}{\pi_i} (i = 1, 2, \dots, N)$ 之间的差别比较小。

如果 $\pi_i > 0, \pi_{ij} > 0 (i, j = 1, 2, \dots, N; i \neq j)$, 则 $V(\hat{Y}_{HT})$ 的无偏估计为:

$$v(\hat{Y}_{HT}) = \sum_{i=1}^n \frac{1 - \pi_i}{\pi_i^2} y_i^2 + 2 \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} y_i y_j \quad (5.12)$$

如果 n 固定, 则 $V(\hat{Y}_{HT})$ 也可用 Yates, Grundy 和 Sen 提出的

$$v_{YGS}(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (5.13)$$

来估计, 它也是无偏估计。

在实际工作中, 这两个估计式都不是特别理想, 因为它们都有可能为负值。但当 $n = 2$ 时, $v_{YGS}(\hat{Y}_{HT})$ 总是大于零。

下面我们不放回不等概抽样来说明不等概抽样要比简单随机抽样的估计精度高。

【例 5.3】 假设有 5 个居委会, 每个居委会的住户数 X 已知, 但常住居民人数未知, 我们从这 5 个居委会抽出 2 个来估计常住居民的总人数。如表 5.3。

表 5.3 5 个居委会有关数据

i	住户数(X)	常住居民人数(Y)	包含概率(π_i)
1	400	1 100	0.8
2	250	600	0.5

续前表

i	住户数(X)	常住居民人数(Y)	包含概率(π_i)
3	200	500	0.4
4	100	240	0.2
5	50	80	0.1
Σ	1 000	2 520	—

表中的包含概率为:

$$\pi_i = n \frac{X_i}{\sum_{i=1}^N X_i} = n \frac{X_i}{X_0}$$

从 5 个居委会中不放回地抽出 2 个居委会, 不论是不放回不等概抽样还是简单随机抽样, 共有 10 种不同的样本, 我们对这些样本分别利用霍维茨 - 汤普森估计量及简单随机抽样简单估计计算对总量的估计。如表 5.4。

表 5.4 不同估计量的估计结果

样 本	$\hat{Y}_{\pi ps}$	\hat{Y}_{sr}
1,2	2 575	4 250
1,3	2 625	4 000
1,4	2 575	3 350
1,5	2 175	2 950
2,3	2 450	2 750
2,4	2 400	2 100
2,5	2 000	1 700
3,4	2 450	1 850
3,5	2 050	1 450
4,5	2 000	800

从理论上说, $\hat{Y}_{\pi ps}$ 和 \hat{Y}_{sr} 都是无偏估计, 它们的均值是 2 520, 为计算估计量的均值, 必须计算每个样本被抽出的概率。对于简单随机抽样, 每个样本被抽出的概率相同, 因此可以对上述 10 个样本的估计进行简单平均, 但不放回不等概样本, 每

个样本被抽出的概率的计算并不容易。

为比较估计量的优劣,需计算估计量的方差,这也用到每个样本被抽出的概率,不过从1.例,我们可以看出, $\hat{Y}_{\pi ps}$ 比 \hat{Y}_{or} 更集中于总体均值 因此,不放回不等概霍维茨 - 汤普森估计量比简单随机抽样简单估计更精确,出现这种结果是因为 X 和 Y 之间有较强的相关关系。

三、 n 不同情况下的严格 π PS 抽样

我们在上面提到的严格的 π PS 抽样,就是指 n 固定、严格不放回、包含概率 π_i 与单元大小严格成比例,即 $\pi_i = nZ_i$,下面分别介绍一种适合于 $n = 2$ 和 $n > 2$ 情形的严格的 π PS 抽样

(一) $n = 2$ 的情形

对于 $n = 2$ 的情形,在总体中只抽2个单元,因此,通常用逐个抽取法来保证抽样是不放回的,我们可以采用几种不同的抽样方法。对总体所有的单元,如果有 $Z_i < \frac{1}{2}$, 就可以采用 Brewer(布鲁尔)方法。

Brewer 方法的两个样本单元的抽取方法是:按与 $\frac{Z_i(1-Z_i)}{1-2Z_i}$ 成比例的概率抽取第一个单元,记第一个被抽出的单元为 j ,按与 $\frac{Z_i}{1-Z_j}$ 成比例的概率在剩下的 $N-1$ 个单元中抽取第二个单元。

Brewer 方法的包含概率为:

$$\pi_i = 2Z_i$$

$$\pi_{ij} = \frac{4Z_i Z_j (1 - Z_i - Z_j)}{(1 - 2Z_i)(1 - 2Z_j) \left(1 + \sum_{l=1}^N \frac{Z_l}{1 - 2Z_l} \right)} \quad (5.14)$$

于是对总体总量估计可采用 Horvitz - Thompson 估计量:

$$\hat{Y}_B = \frac{y_i}{\pi_i} + \frac{y_j}{\pi_j} = \frac{1}{2} \left(\frac{y_i}{z_i} + \frac{y_j}{z_j} \right) \quad (5.15)$$

$$v_{YOS}(\hat{Y}_{HT}) = \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (5.16)$$

【例 5.4】 对于例 5.3,如果抽样是按 Brewer 方法的,则其所有可能样本的 π_{ij} 如表 5.5

表 5.5

所有可能样本的 π_{ij} 计算结果

样 本	$\hat{Y}_{\pi ps}$	π_{ij}
1,2	2 575	0.348 79
1,3	2 625	0.265 74
1,4	2 575	0.124 57
1,5	2 175	0.060 90
2,3	2 450	0.091 35
2,4	2 400	0.040 48
2,5	2 000	0.019 38
3,4	2 450	0.029 07
3,5	2 050	0.013 84
4,5	2 000	0.005 88

于是,可以按下述公式

$$E(\hat{Y}_{\pi ps}) = \sum \hat{Y}_{\pi ps} \pi_{ij}$$

$$V(\hat{Y}_{\pi ps}) = \sum (\hat{Y}_{\pi ps} - Y)^2 \pi_{ij}$$

计算 $\hat{Y}_{\pi ps}$ 的均值及方差,它们分别是 2 520 和 22 670.93。与简单随机抽样简单估计 \hat{Y}_{sr} 的方差 1 151 100 相比, $\hat{Y}_{\pi ps}$ 比 \hat{Y}_{sr} 精确得多。

(二) $n > 2$ 的情形

对于 $n > 2$ 的情形,也有几种不同的抽样方法,例如 Brewer 方法就可以从 $n = 2$ 推广到 $n > 2$ 的情形,但它的 π_{ij} 计算相当复杂。下面介绍一种比较方便的方法——水野法。

水野法也是一种逐个抽取的方法,它以概率

$$Z_i^* = \frac{n(N-1)Z_i}{N-n} = \frac{n-1}{N-n}, \quad i = 1, 2, \dots, N \quad (5.17)$$

抽取第一个样本单元,在剩下的 $N-1$ 个单元中,不放回、等概率地抽出 $n-1$ 个样本单元,为了保证每个 $Z_i^* \geq 0$,要求每个单元的大小满足

$$M_i \geq \frac{(n-1)M_0}{n(N-1)} \quad (5.18)$$

为满足这一点,必须避免 M_i 相差过大,我们可以通过分层,将大小相似的单元分

到同一个层来解决这个问题。

对于水野法,其包含概率为:

$$\pi_i = nZ_i \quad (5.19)$$

$$\pi_{ij} = \frac{n}{N-1} \left[\frac{N-n}{2} (Z_i^* + Z_j^*) + \frac{n-2}{N-n} \right] \quad (5.20)$$

将其代入 Horvitz - Thompson 估计量就可对总体总量进行估计。

四、几种非严格的 π PS 抽样

在实际工作中,我们有时采用非严格的 π PS 抽样,就是指 n 不固定,而是随机的;或不是严格不放回的;或包含概率 π_i 与单元大小并非严格成比例,即 $\pi_i = nZ_i$ 不严格成立。

(一) Yates - Grundy 逐个抽取法

Yates - Grundy(耶茨 格伦迪) 逐个抽取法,每次都以与未入样的单元的 Z_i 成比例的概率抽样,即以 Z_i 抽取第一个样本单元,不妨记被抽中的单元为第 1 个;

以 $\frac{Z_i}{1 - Z_1}$ 在剩下的 $N - 1$ 个单元中抽取第二个样本单元,不妨记被抽中的单元为

第 2 个;以 $\frac{Z_i}{1 - Z_1 - Z_2}$ 在剩下的 $N - 2$ 个单元中抽取第三个样本单元;依此类推,直到抽出 n 个样本单元。这种方法显然 π_i 不是与单元大小严格成比例的,但它在不放回不等概抽样中操作最简单、想法最自然,因而在实际中人们常常使用。

Yates - Grundy 方法的 π_i 不易计算,因而不能用 Horvitz - Thompson 估计量。我们可以采用 Raj(拉奇) 估计量。

设 y_1, y_2, \dots, y_n 为按抽中顺序排列的样本单元的指标值,相应的 Z 值为 z_1, z_2, \dots, z_n , 令

$$\begin{cases} t_1 = \frac{y_1}{z_1} \\ t_2 = y_1 + \frac{y_2}{z_2}(1 - z_1) \\ \vdots \\ t_n = y_1 + y_2 + \dots + y_{n-1} + \frac{y_n}{z_n}(1 - z_1 - z_2 - \dots - z_{n-1}) \end{cases} \quad (5.21)$$

则 Raj 估计量为:

$$\hat{Y}_{Raj} = \frac{1}{n} \sum_{i=1}^n t_i \quad (5.22)$$

它是总体总量 Y 的无偏估计, 对其方差 $V(\hat{Y}_{Raj})$ 的无偏估计为:

$$v(\hat{Y}_{Raj}) = \frac{1}{n(n-1)} \sum_{i=1}^n (t_i - \hat{Y}_{Raj})^2 \quad (5.23)$$

【例5.5】 总体由 $N = 10$ 个单元组成, $M_0 = 30$, 要在总体中按不等概逐个抽取法抽出 $n = 3$ 个样本单元, 并在调查后对总体进行推算。

解: (1) 首先利用代码法来进行抽样。如表 5.6。

表 5.6 抽样过程表述表

i	抽取第一个样本单元 代码 M_i	抽取第二个样本单元 代码 M_i	抽取第三个样本单元 代码 M_i
1	3	3	3
2	1	1	1
3	6	6	6*
4	4*		
5	7	7	7
6	3	3	3
7	1	1	1
8	2	2*	
9	2	2	2
10	1	1	1
Σ	$M_0 = 30$	$M_0 - M_4 = 26$	$M_0 - M_4 - M_8 = 24$

如果在 $[1, 30]$ 的范围内产生的随机数为 12, 则代码 12 所在的 4 号单元被抽中。

如果在 $[1, 26]$ 的范围内产生的随机数为 23, 则代码 23 所在的 8 号单元被抽中。

如果在 $[1, 24]$ 的范围内产生的随机数为 5, 则代码 5 所在的 3 号单元被抽中。

(2) 接下来由样本推算总体。

按被抽出的顺序排列, 样本单元为 4, 8, 3 号单元, 相应的 z_i 值为 $\frac{4}{30}, \frac{2}{30}, \frac{6}{30}$ 。

调查完毕后,如果相应的指标值为 y_4, y_8, y_3 , 先计算

$$t_1 = y_4 + \frac{y_4}{4} - 7.5 y_4$$

$$t_2 = y_4 + \frac{y_8}{z_8} (1 - z_4) = y_4 + \frac{y_8}{2} \left(1 - \frac{4}{30} \right) = y_4 + 13 y_8$$

$$t_3 = y_4 + y_8 + \frac{y_3}{z_3} (1 - z_4 - z_8) = y_4 + y_8 + \frac{y_3}{6} \left(1 - \frac{4}{30} - \frac{2}{30} \right) \\ = y_4 + y_8 + 4 y_3$$

将 $n = 3$ 及 t_1, t_2, t_3 代入

$$\hat{Y}_{Raj} = \frac{1}{n} \sum_{i=1}^n t_i$$

$$v(\hat{Y}_{Raj}) = \frac{1}{n(n-1)} \sum_{i=1}^n (t_i - \hat{Y}_{Raj})^2$$

则得到总体总量的估计及其方差的样本估计。

(二) Poisson 抽样

Poisson(泊松)抽样是一种严格不放回, $\pi_i = nZ_i$ 严格成立, 但样本量 n 事先不能确定的抽样方法, 由 Hajek(哈杰克)设计。

这种方法对总体每个单元赋予一个入样概率 π_i , 即设定一个常数 n_0 , 使得 $\pi_i = n_0 Z_i$ 。然后对总体每个单元分别产生一个 $[0, 1]$ 之间的随机数 r , 如果 $r < \pi_i$, 则这个单元被抽中, 否则, 这个单元就未被抽中。这类似于对每个单元分别以一定的中奖概率进行抽奖, 结果是每个单元都有两种可能, 要么中奖, 要么不中奖。例如, 某个单元入样概率为 0.82, 则产生 00—99 之间的一个随机数 (00 对应 100), 不妨这个随机数为 63, 则 $[0, 1]$ 之间的随机数 $r = 0.63$, 这里 $0.63 < 0.82$, 因此, 这个单元被抽中。

这时对总体总量 Y 的估计可以仍旧采用 Horvitz - Thompson 估计量:

$$\hat{Y}_{PS} = \sum_{i=1}^n \frac{y_i}{\pi_i} \quad (5.24)$$

它是总体总量 Y 的无偏估计, 对其方差 $V(\hat{Y}_{PS})$ 的无偏估计为:

$$v(\hat{Y}_{PS}) = \sum_{i=1}^n (1 - \pi_i) \frac{y_i^2}{\pi_i^2} \quad (5.25)$$

Poisson 法每次的样本量 n 事先不能确定, 一种极端的情形是可能一个单元也没被抽中, 或所有的单元都被抽中, 这是它的主要缺点。当然, 如果出现这种极端的

情形,则重新进行 Poisson 抽样,直到样本不出现上述的极端情形为止。

小 结

本章介绍了不等概抽样方法,它是与简单随机抽样方法平行的一类方法。不等概抽样主要用于总体单元差异非常大,而推算目标是总体总量的情形,同时,它也广泛用于整群抽样时群的规模差异较大、多阶段抽样中初级单元差异较大的情况下对群、初级单元的抽取。

不等概抽样的效率比较高,它能大大地提高估计精度。但使用它的条件是,需要说明总体单元大小(规模)的辅助变量来确定每个单元的入样概率或包含概率,这对抽样和推算过程都是需要的。

不等概抽样按抽样时样本单元是否放回可以分为 PPS 抽样和 π PS 抽样。PPS 抽样操作实施相对简单些,严格的 π PS 抽样在 $n = 2$ 时能够实施,对于 $n > 2$ 的情形则比较复杂,通常这时采用不严格的 π PS 抽样。

本章附录 不等概抽样估计量性质的证明

1. 证明汉森 赫维茨估计的性质。

证明:由于 PPS 抽样是从同一个总体中进行 n 次独立抽样,可以设想抽样是从总体

$$\left\{ \frac{Y_1}{Z_1}, \frac{Y_2}{Z_2}, \dots, \frac{Y_N}{Z_N} \right\}$$

中独立抽取的,单元 $\frac{Y_i}{Z_i}$ 被抽中的概率是 Z_i , 这里 $Z_i = \frac{M_i}{M_0}$ 。这时, $\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i}$ 是所获得样本的平均数。因此

$$E(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^n E\left(\frac{y_i}{z_i}\right) = \sum_{i=1}^N \left(\frac{Y_i}{Z_i} Z_i\right) = Y$$

即 \hat{Y}_{HH} 是无偏的。

\hat{Y}_{HH} 的方差是总体方差的 $\frac{1}{n}$, 而总体方差为:

$$\sum_{i=1}^N Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2$$

且其无偏估计为:

$$\frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{Y}_{HH} \right)^2$$

因而, \hat{Y}_{HH} 的方差为:

$$V(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^n Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2$$

$V(\hat{Y}_{HH})$ 的无偏估计为:

$$v(\hat{Y}_{HH}) = \frac{1}{n} \cdot \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{Y}_{HH} \right)^2$$

习 题

1. 某大型企业集团欲对总部及子公司各部门工作情况抽查, 根据时间要求, 准备抽出 n 个部门进行调查。调查人员从人事部门的计算机里获得了集团全体职工名单, 这份名单注明了每个职工工作的部门。调查人员在计算机上顺序给每位职工编号, 最大号为 N , 并利用计算机分别从 $1 \sim N$ 中产生了 n 个伪随机数, 根据这 n 个随机数所对应的号码, 找到了对应的职工, 于是将这 n 个职工所在的部门记录下来, 然后调查者分别对这些部门进行了调查访问。有人认为: “这不是抽部门, 而是抽职工, 而且抽到某个职工则这个部门的所有 (可以看做抽样框中与之相邻的) 职工都被抽中, 这显然违反了随机的原则, 而且操作费事, 应该直接抽部门。” 对此, 你有何评论?

2. 某个调查人员从总体中抽出了一个样本量为 n 的简单随机样本, 调查开始之前, 他又获得了一份总体单元的详细名单, 这份名单很不错, 除了单元的名录, 还有每个单元的其他相关指标, 因此他在调查每个样本单元的时候注明了它们的其他相关指标。调查完成后, 调查人员发现每个单元的目标量 (y_i) 差异非常大, 但目标量除以某个相关指标 (x_i) 之后, $\frac{y_i}{x_i}$ 差异非常小, 因此, 为了提高估计的精度, 他决定采用下述公式进行推算:

$$\hat{Y} = X \cdot \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}$$

$$v(\hat{Y}) = X^2 \cdot \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i}{x_i} - \frac{\hat{Y}}{X} \right)^2$$

式中, $\lambda = \sum_{i=1}^N \lambda_i$, 这个指标可以从调查人员后来获得的总体单元名单中得到

根据上述公式推算得到的结果, 精度相当高, 该调查人员非常满意。你认为该调查人员的工作有需要改进的地方吗?

3. 比较 PPS 抽样 Hansen-Hurwitz 估计量与简单随机抽样比率估计量的差别

4. 对某个总体的调查, 事先规定了每个单元被抽中的概率, 如下表。试利用代码法抽出一个 $n = 3$ 的 PPS 样本。

i	Z	i	Z
1	0.104	6	0.117
2	0.192	7	0.089
3	0.138	8	0.038
4	0.062	9	0.057
5	0.052	10	0.121

5. 如果上题中被抽中的是 3, 4, 10 号单元, 调查得到它们的值分别是 320, 120, 290, 试估计总体总量并计算估计的标准差。

6. 假设总体大小为 $N = 6$, 说明单元大小的指标值分别为 2, 9, 3, 2, 1, 6, 拟采用 $n = 2$ 的 π PS 抽样, 试列出所有可能的样本, 计算每个单元的人样概率 π_i 及每对单元入样的概率 π_{ij} , 并验证式(5.5) 式(5.6)。

7. 对于下述人为总体:

i	1	2	3	4	5
Y_i	10	9	5	2	4
X_i	7	5	3	1	2

如果 $n = 1$, 试计算下述估计量, 并对各种估计量的均方误差进行比较。

- (1) 简单随机抽样简单估计;
- (2) 简单随机抽样比估计;
- (3) PPS 抽样 Hansen-Hurwitz 估计。



第 6 章

整 群 抽 样

前面几章提到抽取样本单元时,都是指组成总体的基本单元,即抽样单元和基本单元是一致的。由若干有联系的基本单元所组成的集合称为群。抽样时抽取群,并对入选群的所有基本单元进行调查,称这种方法为整群抽样。第一节介绍整群抽样的定义、特点和如何划分群,第二节介绍群大小相等条件下的估计,第三节介绍群大小不等条件下的估计,第四节介绍有关整群抽样中比例估计的问题。

§ 6.1 引 言

一、整群抽样的定义与特点

(一) 定义

整群抽样(cluster sampling)是将总体划分为若干群,然后以群(cluster)为抽样单元,从总体中随机抽取一部分群,对中选群中的所有基本单元进行调查的一种抽样技术。

例如,欲对某校学生进行抽样调查,可以采用两种不同的抽样方法,一种是根据学生名录随机抽取学生,然后对被选中的学生实施调查;另一种方法不是直接抽

选了生,而是随机抽取若干间学生宿舍,然后对住在该宿舍的所有学生实施调查。后一种方法就是整群抽样。由此可知,与前面几章所介绍的抽样方式的不同点在于:在整群抽样中,抽样单元与接受调查的基本单元是不同的。由若干个基本单元所组成的集合称为群,调查时以群为抽样单元抽取样本,然后对样本中所包含的所有基本单元进行调查。

从方法上看,整群抽样是由一阶抽样向多阶段抽样过渡的桥梁。在第一阶段抽样中,如果抽出群后即对其中的所有单元进行调查,是单阶段整群抽样。如果抽出群单元后,进一步从中按低一级的单元抽取子样本(二阶段),即两阶段抽样。也可以进一步在子样本的各单元中按更低一级的单元再抽子样本(三阶段),等等。最后

一个阶段所抽出的单元可以是最终基本单元,也可以仍然是群体(基本单元的集合)。对于前者一般称为多阶段抽样,这部分内容将在下一章讨论。对于后者称为多阶段整群抽样。事实上,多阶段整群抽样是多阶段抽样中的一种情形,故本章仅对单阶段整群抽样进行讨论。

(二) 特点

1. 抽样框编制得以简化。抽样调查中需要有包括所有总体基本单元的抽样框,才能应用前几章所介绍的抽样方式抽取样本。但是在实践中,有时构造这样的抽样框是不可能的,因为没有相应的资料,有时虽然可以构造这样的抽样框,但工作量极大。比较而言,构造群的抽样框则要容易、方便一些。例如对北京市小学生的视力状况进行抽样调查,要获得北京市所有小学生的名单十分困难,但若以学校作为群,得到北京市所有小学校的名单则要容易得多。

2. 实施调查便利,节省费用。在总体基本单元分布很广的情形下,简单随机抽样会使样本分布过于分散,给调查带来不便,并使调查费用增大。而整群抽样调查单元的分布相对集中,调查人员能节省大量来往于调查单元间的时间和费用。而且,如果群是以行政单位划分的,调查时得到行政单位的配合,更有助于调查的实施,可得到较高质量的原始数据。

整群抽样的主要弱点是,通常情况下其抽样误差较大。因为抽取的样本单元比较集中,一个群内各单元之间的差异比较小,而不同群之间的差别比较大,这样每个样本单元所提供的信息价值量就很有限,因此抽样误差常常大于简单随机抽样。但由于整群抽样省时省力,每个单元的平均调查费用较少,故可以通过适当增大样本量的方法弥补估计精度的损失。

但是,对于某些特殊结构的总体,整群抽样反而有较高的精度。这种特殊结构的总体是指,总体中各个群的结构相似,例如一般家庭成员中都有男性、女性,如果估计男女性别比例,以家庭作为群,采用整群抽样,估计的精度要比直接抽取个人

进行估计的精度高。

二、群的划分

整群抽样中的“群”大致可分为两类,一类是根据行政或地域形成的群体,如学校、企业或街道,对此采用整群抽样是为了方便调查、节省费用;另一类群则是调查人员人为确定的,如将一大块面积划分为若干块较小面积的群,这时就需要考虑如何划分群,以使在相同调查费用下抽样误差最小。

分群的一般原则可以用方差分析的原理说明。当总体划分为群以后,总体方差可以分解为群间方差和群内方差两部分,这两部分是此消彼长的关系,若群间方差大则群内方差小;反之,群间方差小则群内方差大。由于整群抽样是对入选群中的所有单元都进行调查,因此影响整群抽样误差大小的主要是群间方差。为了提高整群抽样估计的精度,划分群时就应使群内方差尽可能大,而使群间方差尽可能小。换句话说,划分群时应力争使同一群内各单元之间的差异尽可能大,以避免同一群内各单元提供重复信息。这个原则与分层抽样中划分层的原则恰好相反。由此看来,整群抽样和分层抽样是针对不同总体结构而提出的两种不同抽样方式。当然,对于一些复杂结构的总体,也可以把两种抽样方式结合起来,以发挥各自的特长。

三、群的规模

群的规模是指组成群的单元的数量。在整群抽样中,群的规模具有相当的灵活性,可以大些,也可以小些。群的规模大,估计的精度差但费用省;群的规模小,估计的精度可以提高但费用增大。实践中确定群的规模涉及多种因素,如群的具体结构、精度费用问题、调查实施的组织管理等。在正常情况下,群的规模不宜过大,对于规模很大的群,通常需要采用多阶段抽样。一些学者利用方差函数与费用函数对群的最优规模进行过理论上的讨论。

群的规模又有两种情况,一种是总体中的各个群规模相等;另一种是总体中各个群的规模不等。本章将分别对这两种情况进行讨论。

§ 6.2 群规模相等时的估计

若总体 N 个群中,每个群所包含的单元数 M 相等,则称群规模相等。实际问题中只要群规模接近,也可视为群规模相等,这时,一般采用简单随机抽样抽取群。

一、符号说明

总体群数: N

样本群数: n

总体第 i 群中第 j 个单元的指标值: Y_{ij}

样本第 i 群中第 j 个单元的观测值: y_{ij}

第 i 群中的单元数: M_i 在本节中, 各群规模相等, 故有

$$M_1 = M_2 = \cdots = M_N = M \text{ (各 } M_i \text{ 相等记作 } M \text{)}$$

总体中单元总数: $M_0 = \sum_{i=1}^N M_i$

总体中第 i 群的群总值: $Y_i = \sum_{j=1}^M Y_{ij}$

样本中第 i 群的群总值: $y_i = \sum_{j=1}^M y_{ij}$

总体中第 i 群的个体均值: $\bar{Y}_i = \frac{Y_i}{M}$

样本中第 i 群的个体均值: $\bar{y}_i = \frac{y_i}{M}$

总体中的群均值: $\bar{Y} = \sum_{i=1}^N \frac{Y_i}{N}$

样本中的群均值: $\bar{y} = \sum_{i=1}^n \frac{y_i}{n}$

总体中的个体均值: $\bar{Y} = \frac{Y}{M}$

样本中的个体均值: $\bar{y} = \frac{y}{M}$

总体方差: $S^2 = \frac{1}{M_0 - 1} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y})^2$

总体群间方差: $S_b^2 = \frac{M}{N - 1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2$

总体群内方差: $S_w^2 = \frac{1}{N(M - 1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2$

样本方差: $s^2 = \frac{1}{nM - 1} \sum_{i=1}^n \sum_{j=1}^M (y_{ij} - \bar{y})^2$

样本群间方差: $s_b^2 = \frac{M}{n - 1} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2$

$$\text{样本群内方差: } s_u^2 = \frac{1}{n(M-1)} \sum_{i=1}^n \sum_{j=1}^M (y_{ij} - y_i)^2$$

二、估计量

(一) 均值估计量的定义

采用整群抽样,若群的抽取是简单随机的,且群的大小相同,皆等于 M ,则对总体均值 \bar{Y} 的估计为:

$$\bar{y} = \frac{\sum_{i=1}^n \sum_{j=1}^M y_{ij}}{nM} = \frac{1}{n} \sum_{i=1}^n y_i \quad (6.1)$$

(二) 估计量 \bar{y} 的性质

性质 1 \bar{y} 是 \bar{Y} 的无偏估计。即

$$E(\bar{y}) = \bar{Y} \quad (6.2)$$

这是显然的。因为是按简单随机方法抽取群,因此样本群均值 y 是总体群均值 Y 的无偏估计,因而 $E(\bar{y}) = \frac{Y}{M} = \bar{Y}$ 。

性质 2 \bar{y} 的方差为:

$$V(\bar{y}) = \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 \quad (6.3)$$

$$= \frac{1-f}{nM} S_b^2 \quad (6.4)$$

由前而符号说明知, $y = M\bar{y}$, 又

$$M^2 V(\bar{y}) = V(y) = \frac{1-f}{n} \cdot \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N-1}$$

故

$$\begin{aligned} V(\bar{y}) &= \frac{1-f}{nM^2} \cdot \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N-1} \\ &= \frac{1-f}{nM} \cdot \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{M(N-1)} = \frac{1-f}{nM} S_b^2 \end{aligned}$$

性质 3 $V(\bar{y})$ 的样本估计为:

$$v(\bar{y}) = \frac{1-f}{nM} s_b^2 \quad (6.5)$$

由于 s_h^2 是 S_h^2 的无偏估计, 因而 $v(\bar{y})$ 是 $V(\bar{y})$ 的无偏估计。

总体总值 $Y = NMY$ 的估计量及相应的方差可以根据前面结果直接推出, 即

$$\hat{Y} = NM\bar{y} \quad (6.6)$$

$$V(\hat{Y}) = V(NM\bar{y}) = N^2 M^2 V(\bar{y}) \quad (6.7)$$

$$v(\hat{Y}) = N^2 M^2 v(\bar{y}) \quad (6.8)$$

三、整群抽样效率分析

整群抽样的估计精度与群内相关系数有关。群内相关系数 ρ 描述的是同一群内成对个体单元之间的相关程度, 表达式为:

$$\rho = \frac{E(Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y})}{E(Y_{ij} - \bar{Y})^2} \quad (6.9)$$

式中, 分子是对每个群中 M 个个体单元两两配对的离差乘积求平均, 然后再就 N 个群求平均, 因此这样的离差乘积的个数共有 $NC_M^2 = \frac{NM(M-1)}{2}$ 个。于是式 (6.9) 中的分子为:

$$\frac{\sum_{i=1}^N \sum_{j < k}^M (Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y})}{\frac{NM(M-1)}{2}}$$

式 (6.9) 中的分母是对 NM 个个体单元的离差平方项求平均, 故可以写为:

$$\frac{\sum_{i=1}^N \sum_{j < k}^M (Y_{ij} - \bar{y})^2}{NM} = \frac{NM-1}{MN} S^2$$

于是 ρ 又可以写为:

$$\rho = \frac{2 \sum_{i=1}^N \sum_{j < k}^M (Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y})}{(M-1)(NM-1)S^2} \quad (6.10)$$

事实上, 估计量 \bar{y} 的方差可以用群内相关系数近似表示:

$$\begin{aligned} V(\bar{y}) &= \frac{1}{M^2} V(\bar{y}) \\ &= \frac{1}{nM^2} \cdot \frac{1}{N-1} \sum_{i=1}^N (Y_i - Y)^2 \\ &= \frac{1}{n} \cdot \frac{f}{M^2(N-1)} \cdot S^2 \cdot [1 + (M-1)\rho] \end{aligned}$$

$$\approx \frac{1}{nM} f S^2 [1 + (M-1)\rho] \quad (6.11)$$

若采取简单随机抽样,直接从总体中抽取 nM 个个体单元,则样本均值 \bar{y} 的方差公式为:

$$V_{rs}(\bar{y}) = \frac{1}{nM} f S^2$$

由此,可以计算等群(群规模相等)抽样的设计效应:

$$deff = \frac{V(\bar{y})}{V_{rs}(\bar{y})} \approx 1 + (M-1)\rho \quad (6.12)$$

这说明,整群抽样的方差约为简单随机抽样方差的 $1 + (M-1)\rho$ 倍。也就是说,为了得到相同的估计精度,整群抽样的样本量是简单随机抽样样本量的 $1 + (M-1)\rho$ 倍。

整群抽样的估计效率,与群内相关系数 ρ 关系密切。如果群内各单元的值都相等,则群内方差 $S_u^2 = 0$,此时 $\rho = 1$ 为最大值,在这种情况下 $deff = M$,即整群抽样的估计量方差是简单随机抽样估计量方差的 M 倍;若群内方差与总体方差相等,意味着分群是完全随机的,这时 $\rho \approx 0$, $deff = 1$,整群抽样与简单随机抽样估计效率相同;当群内方差大于总体方差时, ρ 的取值为负,这时 $deff < 1$,整群抽样的效率高于简单随机抽样。当群间方差 $S_b^2 = 0$,即各群均值 \bar{Y}_i 都相等时, ρ 有极小值 $-\frac{1}{M-1}$,所以 ρ 的取值范围是 $\left[-\frac{1}{M-1}, 1\right]$ 。

要提高整群抽样估计的效率,就要通过分群尽可能降低 ρ 值,它是通过增大群内单元之间的差异实现的。这个结论也正是前面所谈及的群的划分原则。当然,对于自然形成的群而言,无法通过调整群内单元而控制 ρ 的取值。这时,要想减少抽样误差,就只能增大样本量。

另外,群内相关系数 ρ 也可以用群内方差 S_u^2 和群间方差 S_b^2 表示,并由样本统计量 s_u^2, s_b^2 估计:

$$\hat{\rho} = \frac{s_b^2 - s_u^2}{s_b^2 + (M-1)s_u^2} \quad (6.13)$$

【例 6.1】 在一次对某寄宿中学在校生零花钱的调查中,以宿舍作为群进行整群抽样。每个宿舍有 6 名学生。用简单随机抽样在全部 315 间宿舍中抽取 $n = 8$ 间宿舍。全部 48 个学生上周每人的零花钱 y_{ij} 及相关计算数据如表 6.1 所示。试估计该学校平均每个学生每周的零花钱 \bar{Y} ,并给出其 95% 的置信区间。

表 6.1

8 个宿舍 48 名学生每周零花钱支出额

单位: 元

	宿舍 1	宿舍 2	宿舍 3	宿舍 4	宿舍 5	宿舍 6	宿舍 7	宿舍 8
学生 1	58	91	123	99	110	111	120	96
学生 2	83	83	89	105	99	100	115	80
学生 3	74	79	94	98	132	116	117	63
学生 4	82	111	109	107	87	99	99	130
学生 5	66	101	79	129	99	107	106	105
学生 6	87	69	80	90	124	105	120	86
\bar{y}	75.00	89.00	95.67	104.67	108.50	106.33	112.83	93.33
s^2	125.60	233.60	299.07	177.87	287.50	42.27	72.57	527.87

解: 已知 $N = 315, n = 8, M = 6, f = \frac{n}{N} = 0.0254$

故

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{75 + 89 + \cdots + 93.33}{8} = 98.17 (\text{元})$$

$$s_b^2 = \frac{M}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$= \frac{6}{8-1} [(75 - 98.17)^2 + \cdots + (93.33 - 98.17)^2] = 928.6648$$

由(6.5)式

$$v(y) = \frac{1-f}{nM} s_b^2 = \frac{1-0.0254}{8 \times 6} \times 928.6648 = 18.8558$$

$$s(\bar{y}) = \sqrt{v(\bar{y})} = \sqrt{18.8558} = 4.3423$$

于是 \bar{y} 置信度为 95% 的置信区间为:

$$98.17 \pm 1.96(4.3423)$$

也即

$$[89.66 \text{ 元}, 106.68 \text{ 元}]$$

【例 6.2】估计例 6.1 中以宿舍为群的群内相关系数与设计效应。

解: 由例 6.1 已计算出样本群间方差 $s_b^2 = 928.6648$

而样本群内方差为:

$$s_w^2 = \frac{1}{n(M-1)} \sum_{i=1}^n \sum_{j=1}^M (y_{ij} - \bar{y}_i)^2$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n \frac{1}{M-1} \sum_{j=1}^{M_i} (y_{ij} - y_i)^2 \\
&= \frac{1}{n} \sum_{i=1}^n s_i^2 \text{ (见表 6.1 最后一行)} \\
&= \frac{1}{8} (125.60 + 233.60 + \cdots + 527.87) \\
&= 220.79
\end{aligned}$$

由(6.13) 式

$$\begin{aligned}
\hat{\rho} &= \frac{s_b^2 - s_w^2}{s_b^2 + (M-1)s_w^2} = \frac{928.6648 - 220.79}{928.6648 + (6-1)220.79} \\
&= 0.348256
\end{aligned}$$

由(6.12) 式

$$\begin{aligned}
deff &= 1 + (M-1)\hat{\rho} \\
&= 1 + (6-1) \times 0.348256 = 2.741
\end{aligned}$$

设计效应 2.741 表明,在这项调查中,为达到同样的估计精度,整群抽样的样本量大约为简单随机抽样样本量的 2.74 倍。若令 n_{rs} 为简单随机抽样的样本量,则

$$n_{rs} = \frac{nM}{deff} = \frac{8 \times 6}{2.74} \approx 18$$

即可达到整群抽样 48 个学生样本量相同的估计精度。

§ 6.3 群规模不等时的估计

采用整群抽样,如果各群规模 M_i 不等,情况会复杂一些。现实中群规模不等的情况更为常见,此时有不同的抽取群的方法和不同的估计方法。本节将对这些方法加以简要讨论。

一、等概抽样,简单估计

此时不考虑群规模不等的影 响,抽样方法与前节群规模相等时相同,估计方法也相同,即采用简单估计。对总体均值 \bar{Y} 的估计为:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} \right) \quad (6.14)$$

可以看出,此公式与前节(6.1) 式相同。

\bar{y} 的方差估计为:

$$v(\bar{y}) = \frac{1-f}{n} \left(\frac{1}{n-1} \right) \sum_{i=1}^n (y_i - \bar{y})^2 \quad (6.15)$$

因为群规模不等,估计时又未考虑权数,所以估计量 \bar{y} 是有偏的,尤其是当各群规模 M_i 差异很大,且 y_i 与 M_i 存在较高相关时,估计量的偏差会很大。该方法的特点是简便,易于掌握和使用。其适用条件是群之间的规模差异相差不很大时的整群抽样估计。

二、等概抽样,加权估计

其基本思路是,以群规模 M_i 为权数,乘以各群均值 y_i ,得到群观察值总和 y_i ,再将样本中 n 个群的群总和平均,求得群总和均值 \bar{y} ,再除以群平均规模 $\bar{M} = \frac{\sum_{i=1}^n M_i}{n}$,求得均值估计。其估计公式为:

$$\bar{y} = \sum_{i=1}^n \frac{M_i y_i}{n\bar{M}} = \frac{1}{n\bar{M}} \sum_{i=1}^n M_i y_i \quad (6.16)$$

$$= \frac{y \cdot N}{M \cdot N} = \frac{\hat{Y}}{M_0} \quad (6.17)$$

如果总体群平均规模 M 未知,可以用样本群平均规模 $\bar{m} = \frac{\sum_{i=1}^n M_i}{n}$ 代替。

由(6.17)式,方便地得到总体总值 Y 的估计:

$$\hat{Y} = M_0 \bar{y} \quad (6.18)$$

式中, $M_0 = \sum_{i=1}^N M_i$ 为总体中的个体单元总数。

估计总体总值 Y ,需要 M_0 ,但是使用整群抽样的原因之一往往是因为没有总体中个体单元的抽样框,但由于总体的群数 N 是已知的,因此可以采用另一个公式:

$$\hat{Y} = N \sum_{i=1}^n y_i \quad (6.19)$$

可以看出,(6.18)式与(6.19)式是等价的。实际上,先利用(6.19)式求 \hat{Y} ,再利用(6.17)式对总体均值进行估计是比较方便的做法。若 M_0 未知,对总体均值进行估计可采用(6.16)式。

上述估计量的方差分别为:

$$V(\hat{Y}) = \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{N-1} \quad (6.20)$$

它的无偏估计为:

$$v(\hat{Y}) = \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} \quad (6.21)$$

对均值估计 \bar{y} 而言:

$$\begin{aligned} V(\bar{Y}) &= \frac{1}{M_0^2} V(\hat{Y}) \\ &= \frac{N^2(1-f)}{M_0^2 n} \cdot \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{N-1} \end{aligned} \quad (6.22)$$

与简单估计相比,此种加权估计的方法考虑了群规模 M_i , 所以估计量 \bar{y} 和 \hat{Y} 分别是 \bar{Y} 和 Y 的无偏估计。但是从方差公式(6.20)和(6.22)看出,估计量的方差与群总值 Y_i 之间的差异有关。如果群规模 M_i 差别很大,通常会造成 Y_i 差异很大。这样,除了估计的无偏性以外,在估计的精度方面,与前种方法相比,并没有明显改观。

三、等概抽样,比率估计

总体均值采用比率估计的形式为:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i} \quad (6.23)$$

与第4章比率估计的区别在于,这里的辅助变量不是 X_i , 而是群的规模 M_i 。从比率估计量的性质可知,它是一个有偏估计。当样本群数 n 很大时,其偏倚很小,故可以忽略不计。

总体总值 Y 的比率估计为:

$$\hat{Y} = M_0 \bar{y} = M_0 \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i} \quad (6.24)$$

根据比率估计量的方差公式,估计量 \bar{y} 与 \hat{Y} 的方差分别是:

$$\begin{aligned} V(\bar{y}) &\approx \frac{1-f}{nM^2} \cdot \frac{\sum_{i=1}^N (Y_i - \bar{Y}M_i)^2}{N-1} \\ &= \frac{1-f}{nM^2} \cdot \frac{\sum_{i=1}^N M_i^2 (Y_i - \bar{Y})^2}{N-1} \end{aligned} \quad (6.25)$$

及

$$\begin{aligned} V(\hat{Y}) &\approx M_0^2 V(\bar{y}) = N^2 M^2 V(\bar{y}) \\ &\approx \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^N (Y_i - \bar{Y}M_i)^2}{N-1} \\ &= \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^N M_i^2 (Y_i - \bar{Y})^2}{N-1} \end{aligned} \quad (6.26)$$

由(6.25)式、(6.26)式看出,估计量 \bar{y} 与 \hat{Y} 的方差取决于群的个体均值 Y_i 的差异。所以,尽管群规模 M_i 差异可能很大,但 \hat{Y}_i 之间的差异却比 Y_i 之间的差异要小得多。因此,与前一种方法相比,在大样本量情况下,比率估计的精度要更高一些。

$V(\bar{y})$ 与 $V(\hat{Y})$ 的样本估计分别为:

$$\begin{aligned} v(\bar{y}) &= \frac{1-f}{nM^2} \cdot \frac{\sum_{i=1}^n (y_i - M_i \bar{y})^2}{n-1} \\ &= \frac{1-f}{nM^2} \cdot \frac{1}{n-1} \left(\sum_{i=1}^n y_i^2 + \bar{y}^2 \sum_{i=1}^n M_i^2 - 2\bar{y} \sum_{i=1}^n M_i y_i \right) \end{aligned} \quad (6.27)$$

及

$$\begin{aligned} v(\hat{Y}) &= \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^n (y_i - M_i \bar{y})^2}{n-1} \\ &= \frac{N^2(1-f)}{n} \cdot \frac{1}{n-1} \left(\sum_{i=1}^n y_i^2 + \bar{y}^2 \sum_{i=1}^n M_i^2 - 2\bar{y} \sum_{i=1}^n M_i y_i \right) \end{aligned} \quad (6.28)$$

四、与群规模成比例的不等概抽样估计

在群规模不等的整群抽样中,如果群规模差异较大,各个群对总体的影响是不同的。这时可以考虑采用不等概方式抽取群。它的好处是,把群的规模作为抽取样本的辅助信息,提高估计的效果,而且方差估计有比较简单的形式。不等概抽样有放回的 PPS 抽样和不放回的 π PS 抽样,其内容已在第 5 章介绍。这里主要以 PPS 抽样为例进行讨论。

群的抽取是按与 M_i 成比例的 PPS 抽样,每次按

$$Z_i = \frac{M_i}{M_0}, i = 1, 2, \dots, N$$

的概率抽取第 i 个群。根据汉森 - 赫维茨估计量,总体总值 Y 的估计为:

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i} = \frac{M_0}{n} \sum_{i=1}^n \frac{y_i}{M_i} \quad M_0 \bar{y} \quad (6.29)$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{M_i} = \frac{1}{n} \sum_{i=1}^n y_i \quad (6.30)$$

由汉森 - 赫维茨估计量的性质知, \hat{Y} 和 \bar{y} 是 Y 和 \bar{Y} 的无偏估计。估计量的方差是:

$$\begin{aligned} V(\hat{Y}) &= \frac{1}{n} \sum_{i=1}^N Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2 \\ &= \frac{M_0}{n} \sum_{i=1}^N M_i (Y_i - \bar{Y})^2 \end{aligned} \quad (6.31)$$

及

$$\begin{aligned} V(\bar{y}) &= \frac{1}{M_0^2} V(\hat{Y}) \\ &= \frac{1}{n M_0} \sum_{i=1}^N M_i (Y_i - \bar{Y})^2 \end{aligned} \quad (6.32)$$

估计量的估计方差则分别为:

$$\begin{aligned} v(\hat{Y}) &= \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{Z_i} - \hat{Y} \right)^2 \\ &= \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2 \\ v(\bar{y}) &= \frac{1}{M_0^2} v(\hat{Y}) \end{aligned} \quad (6.33)$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (6.34)$$

五、方法比较

【例 6.3】 某县有 33 个乡,共 726 个村,该年度某作物总种植面积为 30 525 亩^①。现采用等概抽样随机抽出 10 个乡进行该种作物的产量调查,如表 6.2,要求估计全县总产量,并计算估计量标准差。

表 6.2 10 个乡调查数据

样本乡编号	村庄数 M_i	农作物总产量(乡) y_i (万公斤)	种植面积(乡) x_i (亩)	$y_i - \frac{y}{M_i}$
1	15	22.0	800	1.466 7
2	18	22.8	780	1.266 7
3	26	30.2	1 000	1.161 5
4	14	21.7	700	1.55
5	20	25.3	880	1.265
6	28	31.2	1 100	1.114 3
7	21	26.0	850	1.238 1
8	19	20.5	800	1.079
9	31	33.8	1 200	1.090 3
10	17	23.6	830	1.388 2
合计	209	257.1	8 940	

资料来源:李金昌《抽样调查与推断》,215 页,北京,中国统计出版社,1996。

对此数据,可以采用前述几种方法求解。

(一) 等概抽样,简单估计

由表 6.2 资料,计算平均每个村的产量为:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1.466\ 7 + \cdots + 1.388\ 2}{10} = 1.262(\text{万公斤})$$

进而

$$\hat{Y} = M_0 \bar{y} = 726 \times 1.262 = 916.212(\text{万公斤})$$

① 考虑到抽样调查的具体情况,保留以亩为单位计算,下同。

$$\begin{aligned}
 v(\hat{Y}) &= M_0^2 v(\bar{y}) = \frac{M_0^2(1-f)}{n} \cdot \frac{1}{(n-1)} \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2 \\
 &= \frac{726^2 \left(1 - \frac{10}{33}\right)}{10} \frac{(1.4667 - 1.262)^2 + \cdots + (1.3882 - 1.262)^2}{10-1} \\
 &= 966.19 (\text{万公斤})
 \end{aligned}$$

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} = 31.1 (\text{万公斤})$$

评价:此种方法的估计过程虽不复杂,但却是有偏估计。

(二) 等概抽样,加权估计

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^n y_i = \frac{33}{10} (22.0 + \cdots + 23.6) = 848.43 (\text{万公斤})$$

$$y = \frac{1}{n} \sum_{i=1}^n y_i = 25.71$$

$$\begin{aligned}
 v(\hat{Y}) &= \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^n (y_i - y)^2}{n-1} \\
 &= \frac{33^2(0.697)}{10} = (20.657) = 1567.9 (\text{万公斤})
 \end{aligned}$$

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} = 39.6 (\text{万公斤})$$

评价:此种方法虽可获得无偏估计量,但与前种方法相比,估计量的估计方差没有改观,反而有所增大。该种方法的估计方差与 y_i 之间的差异有关,它适合于 y_i 之间差异不大的整群抽样。

(三) 等概抽样,比率估计

$$\hat{Y} = M_0 \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i} = 726 \cdot \frac{257.1}{209} = 893.08 (\text{万公斤})$$

$$\begin{aligned}
 v(\hat{Y}) &= \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^n (y_i - M_i \bar{y})^2}{n-1} \\
 &= \frac{33^2(0.697)}{10} \times 9.061 = 687.8 (\text{万公斤})
 \end{aligned}$$

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} = 26.2(\text{万公斤})$$

评价:比率估计将群规模 M_i 作为辅助变量引入估计,其估计方差取决于群均值 Y_i 的差异。 Y_i 的差异比 Y_i 的差异要稳定,所以比率估计比前两种方法获得更好的估计效果,但比率估计是有偏估计,当样本群数 n 较大时,比率估计是比较理想的估计方法

进一步分析发现,影响目标变量 Y_i 的因素不仅有村庄数(群规模) M_i ,而且有种植面积 X_i ,而且后者与 Y_i 的关系更为紧密。于是,用种植面积 X_i 作为辅助变量,代替 M_i 的位置进行比率估计,可能会有更好的结果。

(四) 以其他变量为辅助变量的比率估计

已知全县该作物的种植面积总共有 $X = 30\,525$ 亩。采用种植面积为辅助变量的估计结果为:

$$\hat{Y} = X \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = X \hat{R} = 30\,525 \cdot \frac{257.1}{8\,940} = 877.85(\text{万公斤})$$

$$v(\hat{Y}) = \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1}$$

$$= \frac{33^2(0.697)}{10} \cdot \frac{15.1578}{9} = 127.84(\text{万公斤})$$

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} = 11.3(\text{万公斤})$$

评价:与前面几种方法相比,估计量的估计误差最小,估计效果最好。究其原因,作物的乡产量 y_i 不仅与该乡规模(村庄数) M_i 有关,更与该乡的种植面积 x_i 有关。同 $\sum_{i=1}^n (y_i - M_i \bar{Y})^2$ 相比, $\sum_{i=1}^n (y_i - \hat{R}x_i)^2$ 更小,所以,这种方法在本例中不仅优于简单估计和加权估计,也优于以群规模 M_i 为辅助变量的比率估计。进行估计时不涉及群规模的大小,所以既可以用于群规模不等时的估计,也可以用于群规模相等时的估计。使用这种方法的条件是要掌握辅助变量 X 的总体信息,并在调查中能够获取到与目标变量关系密切的辅助变量的资料。

【例 6.4】 某企业欲估计上季度每位职工的平均病假天数。该企业共有 8 个分厂,现用不等概整群抽样拟抽取二个分厂为样本,并以 95% 的置信度计算其置信区间。有关数据及抽样过程如表 6.3

表 6.3

8 个分厂的职工人数资料

分厂编号	职工人数 (M_i)	累积区间
1	1 200	1 ~ 1 200
2	450	1 201 ~ 1 650
3	2 100	1 651 ~ 3 750
4	860	3 751 ~ 4 610
5	2 840	4 611 ~ 7 450
6	1 910	7 451 ~ 9 360
7	390	9 361 ~ 9 750
8	3 200	9 751 ~ 12 950

由于 $n = 3$, 采用 PPS 抽样, 在数字 1 ~ 12 950 之间, 利用随机数表随机抽取 3 个数, 分别是 02 011, 07 972 和 10 281, 于是 3 分厂、6 分厂和 8 分厂入选样本。用 y_1, y_2, y_3 分别表示三个分厂职工的病假天数, 调查结果为: $y_1 = 4\,320, y_2 = 4\,160, y_3 = 5\,790$ 。

估计过程如下:

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ &= \frac{1}{3} \left(\frac{4\,320}{2\,100} + \frac{4\,160}{1\,910} + \frac{5\,790}{3\,200} \right) = 2.02 (\text{天}) \\ v(\bar{y}) &= \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2 = 0.035\,6 (\text{天})\end{aligned}$$

其置信区间为:

$$2.02 \pm 1.96 \sqrt{0.0356} = 2.02 \pm 0.37$$

若估计全企业因病假而损失的人日, 则

$$\hat{Y} = M_0 \bar{y} = 12\,950 \times 2.02 = 26\,159$$

$$v(\hat{Y}) = M_0^2 v(\bar{y}) = 12\,950^2 (0.035\,6) = 5\,970\,209$$

其置信区间为:

$$26\,159 \pm 1.96 \sqrt{5\,970\,209} = 26\,159 \pm 4\,789$$

评价: 对于群规模不等的整群抽样, 采用不等概 PPS 抽样, 可以得到总体目标量的无偏估计, 并且因为估计量具有自加权性质, 从而使估计量和估计量方差都有

比较简明的形式,估计的效率也比较高,是值得优先考虑采用的方法。

此方法的使用条件是,在抽取样本前,要掌握各群规模 M_i 的信息。此外,抽样过程比等概整群抽样略麻烦些。

§ 6.4 总体比例的估计

采用整群抽样估计总体比例时,可以应用前面已介绍过的同样技术。令 A_i 表示第 i 群中具有某种特征的单元数, $p_i = \frac{A_i}{M_i}$ 是具有该种特征的单元数在第 i 群中的比例。按简单随机方法抽取包含 n 群的样本,利用样本信息对总体比例 P 进行估计。

一、群规模相等时的估计

与群规模相等时均值估计的方法相同,因为比例也是均值,即

$$y = \frac{\sum_{i=1}^n y_i}{n} = p \quad y_i = \begin{cases} 1, \text{具有某种性质} \\ 0, \text{其他} \end{cases}$$

由(6.1)式,用 p_i 代替 \bar{y}_i , 有

$$p = \frac{1}{n} \sum_{i=1}^n P_i = \frac{1}{nM} \sum_{i=1}^n A_i \quad (6.35)$$

是总体比例 P 的无偏估计。

式中, p_i 为样本中第 i 群具有某特征单元数的比例; M 为每群中的单元数。

$$V(p) = \frac{1}{n} f \frac{\sum_{i=1}^N (P_i - P)^2}{N-1} \quad (6.36)$$

利用样本资料,可以得到 $V(p)$ 的无偏估计 $v(p)$ 。

$$v(p) = \frac{1}{n(n-1)} f \sum_{i=1}^n (p_i - p)^2 \quad (6.37)$$

二、群规模不等时的估计

若群规模 M_i 不等,仍采用简单随机抽样抽取群,则总体比例的估计量

$$P = \frac{\sum_{i=1}^n A_i}{\sum_{i=1}^n M_i} \quad (6.38)$$

是比率估计的形式。

根据比率估计的性质,其估计量的方差为:

$$\begin{aligned} V(p) &\approx \frac{1-f}{nM^2} \cdot \frac{\sum_{i=1}^n (A_i - PM_i)^2}{N-1} \\ &= \frac{1-f}{nM^2} \cdot \frac{\sum_{i=1}^n M_i^2 (P_i - P)^2}{N-1} \end{aligned} \quad (6.39)$$

式中, $M = \frac{1}{N} \cdot \sum_{i=1}^n M_i$ 为总体中群的平均规模。 $V(p)$ 的估计式为:

$$\begin{aligned} v(p) &= \frac{1-f}{nM^2} \cdot \frac{\sum_{i=1}^n (A_i - pM_i)^2}{n-1} \\ &= \frac{1-f}{nM^2} \cdot \frac{1}{n-1} \left(\sum_{i=1}^n A_i^2 + P^2 \sum_{i=1}^n M_i^2 - 2p \sum_{i=1}^n A_i M_i \right) \end{aligned} \quad (6.40)$$

【例 6.5】 某居民小区有 415 个居民小组,现采用整群等概抽样,随机抽取 25 个小组为样本,调查中的一项内容为估计男、女性别的比例,表 6.4 资料为样本中女性的分布。试以 95% 的置信度估计该小区女性比例的置信区间,并同简单随机抽样方法进行比较。

表 6.4 25 个居民小组总人数及女性人口数

群(i)	居民数(M_i)	女性人数(A_i)	群(i)	居民数(M_i)	女性人数(A_i)
1	8	4	14	10	5
2	12	7	15	9	4
3	4	1	16	3	1
4	5	3	17	6	4
5	6	3	18	5	2
6	6	4	19	5	3
7	7	4	20	4	1

续前表

群(<i>i</i>)	居民数(<i>M_i</i>)	女性人数(<i>A_i</i>)	群(<i>i</i>)	居民数(<i>M_i</i>)	女性人数(<i>A_i</i>)
8	5	2	21	6	3
9	8	3	22	8	3
10	3	2	23	7	4
11	2	1	24	3	0
12	6	3	25	8	3
13	5	2	合计	151	72

资料来源:Schaffer 等 *Elementary survey sampling*, 264, PWS KENT Publishing Company, 1990

解:这是群规模不等的比例估计。由(6.38)式,总体比例的估计为:

$$p = \frac{\sum_{i=1}^n A_i}{\sum_{i=1}^n M_i} = \frac{72}{151} = 0.477$$

由于总体的 *M* 未知,用样本值 $m = \frac{\sum_{i=1}^n M_i}{n} = \frac{151}{25} = 6.04$ 替代。又根据(6.39)式

$$\begin{aligned} v(p) &= \frac{1-f}{nm^2} \cdot \frac{1}{n-1} \left(\sum_{i=1}^n A_i^2 + p^2 \sum_{i=1}^n M_i^2 - 2p \sum_{i=1}^n A_i M_i \right) \\ &= \frac{0.94}{25(6.04)^2} \cdot \frac{12.729}{25-1} = 0.00055 \end{aligned}$$

故置信区间为:

$$0.477 \pm 1.96 \sqrt{0.00055} = 0.477 \pm 0.046$$

如果采用简单随机抽样方法,从该小区中抽取 151 人,假定调查结果与表 6.4 相同,即其中女性人数为 72 人,抽样比 *f* 也假定相同,则估计量的估计方差为:

$$\begin{aligned} v_{sr}(p) &= \frac{1-f}{n-1} pq \\ &= \frac{0.94}{151-1} (0.477)(0.523) \\ &= 0.00156 \end{aligned}$$

于是可以计算出设计效应

$$deff = \frac{v(p)}{v_{sr}(p)} = \frac{0.00055}{0.00156} = 0.353$$

这表明,在此项内容的调查中,整群抽样的估计效果明显地好于简单随机抽样。

若取 $M = 6.04$,还可以进一步计算群内相关系数 ρ 。

由(6.12)式

$$1 + (M - 1)\rho = deff$$

即

$$1 + (6.04 - 1)\rho = 0.353$$

解得

$$\rho = \frac{0.353 - 1}{5.04} = -0.128$$

群内相关系数为负表明群内差异大而群间差异小。有一些变量如性别,如果以家庭户为群,群内的家庭成员有男、有女,存在明显差异,而群与群之间的性别结构则存在很大的相似性,对于这样一些变量进行估计,整群抽样往往会有最好的估计效果。

小 结

本章介绍了整群抽样的理论及不同条件下整群抽样的估计方法。整群抽样有构造抽样框相对简单、样本单元相对集中、节省调查费用等优点。整群抽样的缺点是估计的效率比较低。整群抽样有群相等和不相等的情况。在群不相等时,按与群大小成比例的不等概抽样抽群是值得考虑采用的。在整群抽样中,比率估计可以有效地提高估计的效率。如果有与目标量关系密切的辅助信息可以利用,对于提高整群抽样的估计精度就更有帮助。

本章附录 整群抽样估计量性质的证明

1.(6.11)式 $V(\bar{y}) \approx \frac{1}{nM} f S^2 [1 + (M - 1)\rho]$ 的证明。

证明:

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 = \sum_{i=1}^N \left[\sum_{j=1}^M (Y_{ij} - \bar{Y}) \right]^2$$

$$\begin{aligned} & \sum_{i=1}^N \left[\sum_{j=1}^M (Y_{ij} - \bar{Y})^2 + 2 \sum_{j < k}^M (Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y}) \right] \\ &= (NM - 1)S^2 + (NM - 1)(M - 1)S^2\rho \\ &= (NM - 1)S^2[1 + (M - 1)\rho] \end{aligned}$$

故有

$$\begin{aligned} V(y) &= \frac{1}{M} V(y) \\ &= \frac{1-f}{nM^2} \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 \\ &= \frac{1-f}{n} \frac{(NM-1)}{M^2(N-1)} S^2[1 + (M-1)\rho] \end{aligned}$$

当 NM 很大时, M 相对于 NM 很小, 因而 $NM - 1$ 与 $NM - M$ 相差不多, 故

$$\frac{1-f}{n} \frac{(NM-1)}{M^2(N-1)} S^2[1 + (M-1)\rho] \approx \frac{1-f}{nM} S^2[1 + (M-1)\rho]$$

2.(6.13) 式 $\hat{\rho} = \frac{s_b^2 - s_u^2}{s_b^2 + (M-1)s_u^2}$ 的证明。

证明:

由(6.10) 式

$$\rho = \frac{2 \sum_{i=1}^N \sum_{j < k}^M (Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y})}{(M-1)(NM-1)S^2}$$

因为

$$S_b^2 = \frac{1}{M(N-1)} \sum_{i=1}^N (Y_i - \bar{Y})^2$$

而

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 = (NM-1)S^2[1 + (M-1)\rho]$$

于是

$$\begin{aligned} 1 + (M-1)\rho &= \frac{M(N-1)S_b^2}{(NM-1)S^2} \\ \rho &= \frac{M(N-1)S_b^2 - (NM-1)S^2}{(M-1)(NM-1)S^2} \end{aligned} \quad (1)$$

当 N 很大, 而 M 相对于 NM 很小时, $NM - 1 \approx NM - M$, 则可将上式写为

$$\rho \approx \frac{S_b^2 - S^2}{(M-1)S^2} \quad (2)$$

又因为

$$\begin{aligned}
 (NM - 1)S^2 &= \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y})^2 \\
 &= \sum_{i=1}^N M(Y_i - \bar{Y})^2 + \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - Y_i)^2 \\
 &= (N - 1)S_b^2 + N(M - 1)S_u^2
 \end{aligned} \quad (3)$$

则

$$S_b^2 = \frac{(NM - 1)S^2 - N(M - 1)S_u^2}{(N - 1)}$$

将其分别代入(1),(2)式,便有

$$\rho = 1 - \frac{NMS_u^2}{(NM - 1)S^2} \approx 1 - \frac{S_u^2}{S^2} \quad (4)$$

由(3)式

$$S^2 = \frac{1}{NM - 1} [(N - 1)S_b^2 + N(M - 1)S_u^2]$$

因为 s_b^2 是 S_b^2 的无偏估计, s_u^2 是 S_u^2 的无偏估计,故

$$\begin{aligned}
 \hat{S}^2 &= \frac{1}{NM - 1} [(N - 1)s_b^2 + N(M - 1)s_u^2] \\
 &\approx \frac{1}{M} [s_b^2 + (M - 1)s_u^2] \quad (N \text{ 很大时})
 \end{aligned} \quad (5)$$

将(5)式代入(4)式,得

$$\hat{\rho} = \frac{s_b^2 - s_u^2}{s_b^2 + (M - 1)s_u^2}$$

习 题

1. 若欲调查城市的猪肉人均消费量,讨论下列情况下采用街道作为群的整群抽样是否合适,如果不合适,你认为采用什么抽样方式好。

- (1) 少数民族的居住比较集中;
- (2) 少数民族比较均匀地分布在各条街道;
- (3) 少数民族分散在各街道,但比重不同。

2. 带锯厂负责对它的用户进行修理,其修理费用每季结算一次。该厂共有 96 家用户,各拥有不同带锯数,现采用等概简单随机方法抽取 20 家为样本,资料如下:

工厂	锯数	修理费用	工厂	锯数	修理费用
1	3	50	11	8	140
2	7	110	12	6	130
3	11	230	13	3	70
4	9	140	14	2	50
5	2	60	15	1	10
6	12	280	16	4	60
7	14	240	17	12	280
8	3	45	18	6	150
9	5	60	19	5	110
10	9	230	20	8	120

- (1) 估计每一带锯的平均修理费用及置信区间($\alpha = 0.05$);
- (2) 根据上述资料估计 96 家用户总的修理费用及置信区间($\alpha = 0.05$);
- (3) 若已知这 96 家用户有 710 条带锯,利用这一补充信息估计总的修理费用和置信区间;

(4) 若欲估计下一季度的每带锯平均修理费用,绝对误差 $\Delta = 2$,试问应抽取多少户作样本。

3. 邮局欲估计每个家庭的平均订报份数,该辖区共有 4 000 户,划分为 400 个群,每群 10 户,现随机抽取 4 个群,取得资料如下表所示。

群	各户订报数(y_{ij})	y_i
1	1,2,1,3,3,2,1,4,1,1	19
2	1,3,2,2,3,1,4,1,1,2	20
3	2,1,1,1,1,3,2,1,3,1	16
4	1,1,3,2,1,5,1,2,3,1	20

试估计平均每户家庭的订报份数及总的订报份数及估计量的方差。

4. 汽车运输公司抽样检查在使用的车辆中不安全轮胎的比例,在 175 辆车中抽了 25 辆,其不安全的轮胎数如下:

不安全轮胎个数	汽车数
0	5
1	8
2	7
3	2
4	3

要求估计该运输公司的汽车中不安全轮胎的比例及估计量的方差。

5. 某工业系统准备实行一项改革措施。该系统共有 87 个单元, 现采用整群抽样, 用简单随机抽样抽取 15 个单元作样本, 征求入选单元中每个工人对改革措施的意见, 结果如下:

单 元	总人数	赞成人数
1	51	42
2	62	53
3	49	40
4	73	45
5	101	63
6	48	31
7	65	38
8	49	30
9	73	54
10	61	45
11	58	51
12	52	29
13	65	46
14	49	37
15	55	42

(1) 估计该系统同意这一改革人数的比例, 并计算估计标准误;

(2) 在调查的基础上对方案作了修改, 拟再一次征求意见, 要求估计比例的绝对误差不超过 8%, 则应抽取多少个单元作样本。

6. 某集团的财务处共有 48 个抽屉, 里面装有各种费用支出的票据。财务人员欲估计办公费用支出的数额, 随机抽取了其中的 10 个抽屉, 经过清点, 整理出办公费用的票据, 得到下表资料:

抽屉编号	票据数(M_i)	费用额(y_i , 百元)
1	42	83
2	27	62
3	38	45
4	63	112
5	72	96
6	12	58
7	24	75
8	14	58
9	32	67
10	41	80

要求以 95% 的置信度估计该集团办公费用总支出额的置信区间。

7. 为了便于管理,将某林区划分为 386 个小区域。现采用简单随机抽样方法,从中抽出 20 个小区域,测量树的高度,得到如下资料:

区域编号	树木株数 (M_i)	平均高度 (y_i , 尺)	区域编号	树木株数 (M_i)	平均高度 (y_i , 尺)
1	42	6.2	11	60	6.3
2	51	5.8	12	52	6.7
3	49	6.7	13	61	5.9
4	55	4.9	14	49	6.1
5	47	5.2	15	57	6.0
6	58	6.9	16	63	4.9
7	43	4.3	17	45	5.3
8	59	5.2	18	46	6.7
9	48	5.7	19	62	6.1
10	41	6.1	20	58	7.0

估计整个林区树的平均高度及 95% 的置信区间。

8. 某市建筑行业集团共有 48 个单元,有载货汽车 186 辆。按每个单元的车辆拥有量成比例的概率进行放回的 PPS 抽样,共抽取 10 次。对抽中单元的所有车辆调查季度运量(单元:吨)。样本数如下表所示(其中有一单元被抽中 2 次,即 $i = 3, 7$)。试估计全集团的季度总运量及 95% 的置信区间。

单元编号	车辆数(M_i)	单元运量总和(y_i)	平均每车运量(y_i)
1	5	14 230	2 846
2	8	21 336	2 667
3	5	13 650	2 730
4	4	11 568	2 892
5	6	15 216	2 536
6	9	23 049	2 566
7	5	13 650	2 730
8	3	7 443	2 481
9	7	16 723	2 389
10	3	8 391	2 797

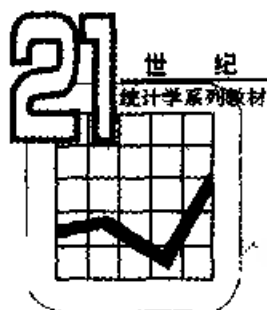
9. 在一次农村调查中,抽样单元是包含 M 个农场的群,当抽取 n 个群作样本时,其费用是 $C = 4tMn + 60\sqrt{n}$,其中 t 是调查一个农场所花的时间(按小时计

算)。如果这一调查的总费用是2 000元,当 $M = 1, M = 5, M = 10, t = 0.5, t = 2$ 时, n 的数值计算如下:

t	M		
	1	5	10
0.5	400	131	74
2	156	40	21

样本均值的方差是 $\frac{S^2}{Mn}[1 + (M-1)\rho]$,有限总体修正系数 fpc 忽略不计。如果 M 在1~10之间, $\rho = 0.1$,试问当(1) $t = 0.5$ 小时,(2) $t = 2$ 小时时, M 多大能得到最精确的结果?怎样解释两个结果的差别?

10. 如果调查经费从2 000元增加到5 000元,你认为原来最优的 M 是增大还是减小,请说明理由。



第 7 章

系 统 抽 样

在实际工作中,系统抽样是一种被广泛采用的抽样方法。系统抽样比简单随机抽样易于操作,但抽样误差的估计比较复杂。实践中,大大小小的抽样调查,尤其是大规模抽样调查,如城乡居民住户抽样调查、人口抽样调查、农产量抽样调查、产品质量抽样检查等,都普遍采用系统抽样。本章第一节介绍系统抽样的定义、作用和特点,第二节介绍系统抽样主要方法,第三节介绍等概率系统抽样的估计量,第四节介绍不同特征总体的系统抽样,第五节介绍系统抽样的方差估计。

§ 7.1 引 言

一、定义

系统抽样(systematic sampling)是将 N 个总体单元按一定顺序排列,先随机抽取一个单元作为样本的第一个单元,即起始单元,然后按某种确定的规则抽取其他样本单元的一种抽样方法。系统抽样中最简单也是最常用的规则是等间隔抽取,这种系统抽样又称等距抽样。由于这种抽样方法看来似乎很“机械”,所以系统抽样有时也称为机械抽样。另外,由于系统抽样提供了区别于简单随机抽样的另一个随机

且独立的挑选样本单元的方式,有时也被称为伪随机抽样。

系统抽样的实际应用非常广泛,例如工业企业为检查产品质量,在连续生产线上每隔2小时抽选一个或若干样品进行检验;农作物产量实测或对农作物害虫进行调查,对一大片农田每隔一定距离(例如2平方米)抽取一小块进行测量或调查;图书馆对图书借阅情况进行调查,在一堆按书名字母排列的图书目录卡片中,每隔一定厚度(例如1厘米)或一定的张数抽取一张卡片等等,都是系统抽样的直观案例。

二、系统抽样的一般方法

(一) 直线等距抽样

假设总体单元数为 N , 样本容量为 n , N 是 n 的整数倍。

首先计算抽样间距 $k = \frac{N}{n}$, 把总体分为 n 段, 每段 k 个单元。然后, 在第一段的 k 个单元中随机抽出一个单元, 假设为 r , 然后每隔 k 个单元抽出一个单元, 即 $r + k, r + 2k, \dots$, 直到抽出 n 个单元。抽出的样本是由以下编号的单元组成: $r + (j - 1)k (j = 1, 2, \dots, n)$ 。如图 7.1。

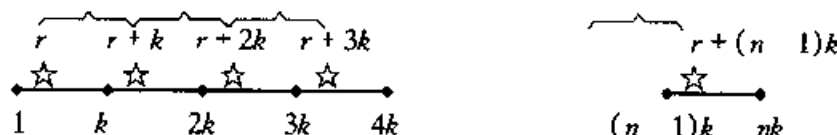


图 7.1 一般直线等距抽样

例如某学院共有 200 个学生, 要抽 10 个学生做样本。首先计算抽样间距 $k = \frac{N}{n} = \frac{200}{10} = 20$, 然后在 1 ~ 20 中随机抽取一个数字, 假设抽中排在第 3 位的学生, 则其余样本单元依次为第 23, 43, 63, 83, 103, 123, 143, 163, 183 位学生。

(二) 循环等距抽样

当 N 不是 n 的整数倍, 即抽样间距 $k = \frac{N}{n}$ 不是整数时, 实际抽取的样本量是不固定的 (k 只能取一个与 $\frac{N}{n}$ 最为接近的整数), 每个总体单元入样的概率也是不等的, 这时用直线等距抽样就有可能产生偏倚。为了使样本均值为无偏估计, 可以采用循环等距抽样方法。其方法是将 N 个总体单元排成首尾相接的一个圆。抽样间距 k 取最接近 $\frac{N}{n}$ 的整数, 从 1 到 N 中随机抽取一个随机起点作为起始单元, 然后每隔 k 抽取一个单元, 直到抽满 n 个单元为止。

例如总体有 14 个单位, 拟抽取 $n = 3$, 则 $k = \frac{N}{n} = 4.7$, 取与之最近的整数 $k = 5$ 。然后在总体中随机抽取一个单位作为起点, 假设抽中 3, 即 $r = 3$, 依次抽取 $r = 3, r + k = 8, r + 2k = 13$, 直到抽满。于是, 样本单位的顺序编号分别为 3, 8, 13。抽样过程见图 7.2。

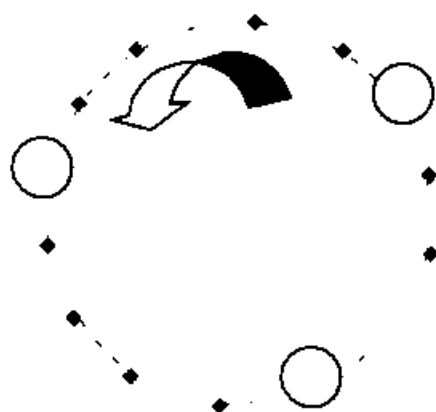


图 7.2 循环等距抽样

(三) 不等概系统抽样法

不等概系统抽样中每个单元的入样概率不相等。最常用也是最简单的不等概率系统抽样 π PS 是系统抽样, 即入样概率 π_i 与单元大小 M_i 成比例的系统抽样。令

$M_0 = \sum_{i=1}^N M_i$, 表示总体所有单元大小的总和, 则

$$\pi_i = n \frac{M_i}{M_0}$$

在实际中, 实施不等概率抽样最简单的方法是代码法。 π PS 系统抽样如下:

先将单元 M_i 值累加, 取最接近 $\frac{M_0}{n}$ 的整数 k 为抽样间距, 从 $[1, k]$ 中随机抽取一个整数 r , 则代码 $r, r + k, \dots, r + (n - 1)k$ 所对应的单元即为样本单元。

【例 7.1】 设总体由 10 个行政村组成, $N = 10$, 每个行政村的人数 M_i 见表 7.1。利用 π PS 系统抽样抽取 $n = 3$ 个行政村。

表 7.1 用 π PS 系统抽样抽选行政村

行政村编号	人数 (M_i)	累计人数	抽中代码
1	103	103	100
2	432	535	
3	96	631	

续前表

行政村编号	人数(M_i)	累计人数	抽中代码
4	246	877	723 1 346
5	84	961	
6	73	1 034	
7	205	1 239	
8	168	1 407	
9	146	1 553	
10	317	1 870	

$$M_0 = \sum_{i=1}^N M_i = 1\,870, n = 3, k = \frac{M_0}{n} = 623$$

从 $[1, k]$ 中随机抽取一个整数 $r = 100$, 则代码 $r = 100, r + 5 = 723, r + 2k = 1\,346$ 所对应的行政村入样, 其序号依次为1, 4, 8。

在 π PS系统抽样中, 对于特别大的单元一定要注意。如果出现 $M_i > k$, 该单元肯定被抽入样本, 而且还可能被重复抽到。为避免这种情况, 可以事先将这些单元从抽样框中提出直接放入样本, 再对由剩余单元组成的总体实施抽样。

三、总体单元的排序

系统抽样时 N 个总体单元的排序情况大致有以下三种。

1. 按无关标志排队。即各单元的排列顺序与所研究的内容无关。例如调查学生的视力状况, 将学生按其学号排序, 学号与视力之间没有必然联系; 又如调查某厂职工平均年龄, 按职工的姓氏笔划排序等。这种排队抽样类似于简单随机抽样, 也称为无序系统抽样。

2. 按有关标志排队。即各单元的排列顺序与所研究的内容是有关系的。例如调查学生的身高, 将全部学生按入校体检时的身高由高到低排队; 又如对农产量进行抽样调查, 将各地块按当年估产或前几年的平均实产由低到高排队。这种排队抽样称为有序系统抽样, 可以使抽取的样本单元更具有代表性, 减小抽样误差, 提高估计的效率。

3. 处于上述两者之间, 根据各单元原有的自然位置进行排序。例如入户调查根据街道门牌号码按一定间隔抽取; 工业生产质量检验每隔一定时间抽取生产线上的产品; 工厂中的工人名单按原有的工资名册顺序等。这种自然状态的排列有时

与调查标识有一定的联系,但又不完全一致,这主要是为了抽样方便。

四、系统抽样的优缺点

作为实践中最常用的抽样方法之一,系统抽样的特点显著,优点和缺点同样明显

(一) 系统抽样的优点

系统抽样的最大优点是简便易行,简化抽样手续。具体来说,系统抽样的优点主要体现在以下两个方面。

1. 简便易行,容易确定样本单元。其他概率抽样方法在抽取样本之前需要对总体单元编号,然后才能利用随机数表等方法抽取样本。当总体单元很多时,编号与抽选都比较麻烦。而系统抽样所需要的只是总体单元的顺序排列,只要随机确定一个(或少数几个)起始单元,整个样本就自然确定,在某些场合下甚至不需要抽样框。例如对公路旁的树木进行病虫害调查,确定每30棵树检查一棵,只要确定了起点的被检树,每隔30棵检查一棵即可,根本不需要事先对路旁的所有树木编号。又如对某市的机动车辆进行调查,确定抽样比为1%,则可在00~99中随机抽取一个整数(如63),只要对车辆牌照号末两位为63的车辆都进行调查即可。

系统抽样不仅实施简单,容易为不熟悉抽样的非专业人员所掌握,而且还因其较易保留抽样过程的原始记录,便于监督和检查,因此在一些大规模抽样调查中,如在多阶段抽样的最后一阶段或二阶段抽样中,经常采用系统抽样以代替简单随机抽样。普查工作中也可以配合使用系统抽样,美国、日本、印度等国都曾从普查资料中系统抽取样本再进行深入调查。

2. 样本单元在总体中分布比较均匀,有利于提高估计精度。如果调查者对总体的结构有一定了解,可以利用已有信息对总体单元进行排列,即按有关标志对总体单元排序,这样采用有序系统抽样就可以有效地提高估计的精度。

(二) 系统抽样的缺点

系统抽样也有其突出的局限性,具体表现为以下两点。

1. 如果单元的排列存在周期性的变化,而抽样者对此缺乏了解或缺乏处理的经验,抽取出样本的代表性就可能很差。例如,商店销售额存在明显的周期性变化,如果系统抽样的样本单元间隔正好与周期变化的长度吻合,不采用一些处理方法进行调整,系统抽样的样本代表性就很差。

2. 系统抽样的方差估计较为复杂,一般系统抽样没有设计意义下的无偏估计量,并且在很多实际应用中所采用的系统抽样都不是严格的概率抽样,这就给系统抽样方差的估计带来很大的困难。

五、系统抽样、整群抽样和分层抽样的关系

在系统抽样过程中，一旦起始单元确定，整个样本就确定了，这是系统抽样有别于其他抽样方法的一大特点。系统抽样既看成一种特殊的整群抽样，又可以看成一种特殊的分层抽样。

以一般的等距抽样为例，假设抽样间距为 k ，总体单元数为 $N = nk$ ，将总体的 N 个单元排列成 k 行 n 列，如表 7.2。显然，表中的每一行单元都是系统抽样的一个样本。

表 7.2 系统抽样的总体单元

	1	2	...	j	...	n	平均
1	Y_1	Y_{k+1}	...	$Y_{(j-1)k+1}$...	$Y_{(n-1)k+1}$	\bar{y}_1
2	Y_2	Y_{k+2}	...	$Y_{(j-1)k+2}$...	$Y_{(n-1)k+2}$	\bar{y}_2
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
r	Y_r	Y_{k+r}	...	$Y_{(j-1)k+r}$...	$Y_{(n-1)k+r}$	\bar{y}_r
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
k	Y_k	Y_{2k}	...	Y_{jk}	...	Y_{nk}	\bar{y}_k

为方便起见，我们按照行列号将总体单元重新编号，令 $Y_{rj} = Y_{(j-1)k+r}$ ($r = 1, 2, \dots, k; j = 1, 2, \dots, n$)，结果见表 7.3。

表 7.3 系统抽样的总体单元按行列重新编号

	1	2	...	j	...	n	群平均
1	Y_{11}	Y_{12}	...	Y_{1j}	...	Y_{1n}	\bar{Y}_1
2	Y_{21}	Y_{22}	...	Y_{2j}	...	Y_{2n}	\bar{Y}_2
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
r	Y_{r1}	Y_{r2}	...	Y_{rj}	...	Y_{rn}	\bar{Y}_r
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
k	Y_{k1}	Y_{k2}	...	Y_{kj}	...	Y_{kn}	\bar{Y}_k
层平均	\bar{Y}_1	\bar{Y}_2	...	\bar{Y}_j	...	\bar{Y}_n	\bar{Y}

如果将每一行单元视为一个群，则总体由 k 个群组成，每个群的大小都是 n 。系统抽样就是从 $Y_{11} \sim Y_{k1}$ 中任选一个单元，被选中单元所在行的所有单元就构

成系统抽样的一个样本,显然,每个群都是一个可能样本,这 k 个可能样本被抽中的概率都等于 $\frac{1}{k}$ 。由于起始单元 $Y_{r,1}$ 都是随机抽取的,因此系统抽样可以看成从 k 个群中随机抽取 1 个群的整群抽样。

同样,将每一列单元视为一层,则总体由 n 个层组成,每个层的大小都是 k ;系统抽样就是从第一层 ($Y_{1,1} \sim Y_{k,1}$) 中任选一个单元,则后面各层中相同行号 (r) 的单元都进入样本。系统抽样可视为从每个层中取一个单元,因此是一种分层抽样,但是由于样本单元在各层的位置相同,因此系统抽样不同于分层随机抽样。

§ 7.2 等概率系统抽样估计量

本节先讨论最简单的系统抽样的估计,即直线等距抽样时总体均值 Y 的估计问题。为方便讨论,假设 $N = nk$,这时抽样是一种严格意义上的概率抽样。

一、符号说明

第 r 行第 j 列的单元指标值: Y_{rj}

$$Y_{rj} = Y_{(r-1)k+j}, r = 1, 2, \dots, k; j = 1, 2, \dots, n$$

总体单元数: N

样本单元数: n

$$\text{系统样本平均数: } \bar{y}_r = \frac{1}{n} \sum_{j=1}^n y_{rj}$$

系统样本均值估计量: y_{ry}

层均值: $y_j, j = 1, 2, \dots, n$

总体方差: S^2

$$\text{系统样本(群)内方差: } S_{usy}^2 = \frac{1}{k(n-1)} \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - \bar{y}_r)^2$$

样本(群)内相关系数: ρ_{usy}

$$\rho_{usy} = \frac{E(y_{rj} - Y)(y_{ru} - Y)}{E(y_{rj} - Y)^2}$$

层内方差: S_{usj}^2

$$S_{usj}^2 = \frac{1}{n(k-1)} \sum_{j=1}^n \sum_{r=1}^k (y_{rj} - y_{\cdot j})^2$$

同一系统样本内对层均值离差的相关系数: ρ_{ust}

$$\rho_{ust} = \frac{E(y_{rj} - y_{\cdot j})(y_{ru} - y_{\cdot u})}{E(y_{rj} - y_{\cdot j})^2}$$

二、估计量

假设起始值为 r , 则相应系统样本的平均数为:

$$y_r = \frac{1}{n} \sum_{j=1}^n y_{rj} = \frac{1}{n} \sum_{j=1}^n Y_{rj} \quad (7.1)$$

取系统样本的平均数作为总体均值 Y 的估计量:

$$y_{sy} = y_r = \frac{1}{n} \sum_{j=1}^n y_{rj} \quad (7.2)$$

性质 1 当 $N = nk$ 时, 有 k 个可能样本:

$$E(y_{sy}) = \frac{1}{k} \sum_{r=1}^k y_r = \frac{1}{nk} \sum_{r=1}^k \sum_{j=1}^n y_{rj} = Y \quad (7.3)$$

因此 y_{sy} 是无偏估计量。

但是当 $N \neq nk$ 时, 采用直线等距抽样得到的 k 个可能样本所包含的单元数不全相等, 因此 y_{sy} 是有偏的。不过, 当 N 和 n 均比较大时, 其偏倚不会很大, 可以忽略不计。如果采用循环等距抽样, y_{sy} 是无偏的。

三、估计量方差的不同表示形式

为方便起见, 以后均假定 $N = nk$ 时, 系统样本的平均数 y_{sy} 作为总体均值的估计是无偏的。它的方差按定义为:

$$V(y_{sy}) = E(y_{sy} - Y)^2 = \frac{1}{k} \sum_{r=1}^k (y_r - Y)^2 \quad (7.4)$$

下面给出方差的三种不同的表示形式。

性质 2 用样本(群)内方差 S_{usy}^2 表示系统抽样估计量的方差:

$$V(y_{sy}) = \frac{(N-1)}{N} S^2 = \frac{k(n-1)}{N} S_{usy}^2 \quad (7.5)$$

式中, $S^2 = \frac{1}{N-1} \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - Y)^2$ 为总体方差; $S_{usy}^2 = \frac{1}{k(n-1)} \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - y_r)^2$ 为样本(群)内方差。

如果从总体 N 中直接抽取样本量为 n 的简单随机样本, 则总体均值 Y 的估计量 y_{rsr} 的方差

$$V(y_{sys}) = \frac{N-n}{Nn} S^2 = \frac{1-f}{n} S^2$$

式中, S^2 为总体方差; n 为样本量; f 为抽样比。

比较等距抽样方差 $V(y_{sys})$ 和简单随机抽样方差 $V(y_{srs})$, 可得出以下结论:

$$\text{对于 } V(y_{srs}) - V(y_{sys}) = \frac{n}{n} \frac{1}{n} (S_{usy}^2 - S^2)$$

有 $\left\{ \begin{array}{l} \text{当 } S_{usy}^2 > S^2, \text{即等距样本内方差大于总体方差时, 系统抽样法优} \\ \quad \text{于简单随机抽样;} \\ \text{当 } S_{usy}^2 < S^2, \text{即等距样本内方差小于总体方差时, 简单随机抽样} \\ \quad \text{优于系统抽样法;} \\ \text{当 } S_{usy}^2 = S^2, \text{即等距样本内方差等于总体方差时, 系统抽样法与} \\ \quad \text{简单随机抽样法抽样效果相同。} \end{array} \right.$

对于固定总体, 总体方差是惟一确定的, 因此, 系统样本内的方差 S_{usy}^2 越大, 系统抽样的精度越高。为了提高系统抽样的精度, 总体单元的排列应尽可能增大样本(群)内方差。

性质 3 系统抽样可看做一种特殊的整群抽样, 系统抽样估计量的方差可以用样本(群)内相关系数 ρ_{usy} 表示:

$$V(y_{sys}) = \frac{S^2}{n} \left(\frac{N-1}{N} \right) [1 + (n-1)\rho_{usy}] \quad (7.6)$$

式中, ρ_{usy} 为样本(群)内相关系数。

$$\begin{aligned} \rho_{usy} &= \frac{E(y_{rj} - Y)(y_{ru} - Y)}{E(y_{rj} - Y)^2} \\ &= \frac{2}{(n-1)(N-1)S^2} \sum_{r=1}^k \sum_{j < u}^n (y_{rj} - Y)(y_{ru} - Y) \end{aligned} \quad (7.7)$$

由性质 3 可见, 系统样本(群)内正相关越大, 即系统样本(群)内单元越相似, 则估计量方差越大, 等距抽样精度越差。该结论与性质 2 的结论显然是是一致的。

性质 4 系统抽样可看做一种特殊的分层抽样, 系统抽样估计量的方差可以用层内方差 S_{ust}^2 和 ρ_{ust} 表示:

$$V(y_{sys}) = \frac{S_{ust}^2}{n} \left(\frac{N-n}{N} \right) [1 + (n-1)\rho_{ust}] \quad (7.8)$$

式中, S_{ust}^2 为层内方差, $S_{ust}^2 = \frac{1}{n(k-1)} \sum_{j=1}^k \sum_{r=1}^n (y_{rj} - y_{\cdot j})^2$; $y_{\cdot j}$ 为层均值($j = 1, 2, \dots, n$); ρ_{ust} 为同一系统样本内对层均值离差的相关系数。

$$\rho_{ust} = \frac{E(y_{rj} - y_{\cdot j})(y_{ru} - y_{\cdot u})}{E(y_{rj} - y_{\cdot j})^2}$$

$$= \frac{2}{n(n-1)(k-1)S_{wst}^2} \sum_{r=1}^k \sum_{j < u}^n (y_{rj} - y_{\cdot j})(y_{ru} - y_{\cdot u}) \quad (7.9)$$

比较系统抽样方差 $V(y_{sy})$ 与比例分配的分层随机抽样方差 $V(y_{st})$, 比例分配的分层随机抽样总体均值估计量的方差

$$V(y_{st}) = \frac{S_{wst}^2}{n} \left(\frac{N-n}{N} \right)$$

$$\frac{V(y_{sy})}{V(y_{st})} = 1 + (n-1)\rho_{wst}$$

因此

- 当 $\rho_{wst} > 0$ 时, 系统抽样的精度低于分层随机抽样;
- 当 $\rho_{wst} = 0$ 时, 系统抽样的精度与各层随机抽取一个单位的分层随机抽样相同;
- 当 $\rho_{wst} < 0$ 时, 系统抽样的精度高于分层随机抽样。

【例 7.2】 设某个总体有 $N = 32$ 个单元, 总体单元排列显然有稳定上升的趋势。我们要在产生一个样本量为 4 的等距样本, 将总体单元排列如表 7.4, $k = 8$, $n = 4$, 每一列都是一个等距样本, 共 8 个等距样本。

表 7.4 $N = 32, k = 8, n = 4$ 等距样本数据

层	等距样本编号								层均值
	1	2	3	4	5	6	7	8	
I	1	1	3	3	4	5	6	7	3.75
II	7	8	8	11	12	14	16	16	11.5
III	17	18	20	20	24	24	25	27	21.875
IV	27	28	30	31	34	34	36	38	32.25
总数	52	55	61	65	74	77	83	88	—

显然, 层内有正相关, 前 4 个样本与各层均值的离差都是正数, 后 4 个样本与各层均值的离差都是负数, 由性质 4, 当 $\rho_{wst} > 0$ 时, 系统抽样的精度低于分层随机抽样。

层内方差与总方差分别为:

$$S_{wst}^2 = \frac{1}{n(k-1)} \sum_{j=1}^k \sum_{r=1}^n (y_{rj} - \bar{y}_{\cdot j})^2 = 11.5$$

$$S^2 = \frac{1}{N-1} \sum_{j=1}^k \sum_{r=1}^n (y_{rj} - \bar{Y})^2 = 129.523$$

因此,简单随机抽样均值估计的方差 $V(y_{srs})$ 、分层随机抽样均值估计的方差 $V(y_{st})$ 以及等距抽样均值估计的方差 $V(y_{sy})$ 如下:

$$V(y_{srs}) = E(y_{srs} - Y)^2 = \frac{1}{k} \sum_{i=1}^k (y_i - Y)^2 = 9.452$$

$$V(y_{st}) = \frac{S_{ust}^2}{n} \left(\frac{N-n}{N} \right) = \frac{11.5}{4} \times \frac{32-4}{32} = 2.516$$

$$V(y_{sy}) = \frac{N-n}{Nn} S^2 = \frac{32-4}{32} \times \frac{129.5232}{4} = 28.333$$

本例中,分层随机抽样和等距抽样都比简单随机抽样更有效,而分层随机抽样比等距抽样更有效。

【例 7.3】 利用例 7.2 的数据,但将第二层和第四层的观测值次序颠倒,数据见表 7.5。

表 7.5 第二层和第四层的观测值次序颠倒后的等距样本数据

层	等距样本编号								层均值
	1	2	3	4	5	6	7	8	
I	1	1	3	3	4	5	6	7	3.75
II	16	16	14	12	11	8	8	7	11.5
III	17	18	20	20	24	24	25	27	21.875
IV	38	36	34	34	31	30	28	27	32.25
总数	72	71	71	69	70	67	67	68	—

显然,等距样本内数据与各层均值的离差有正有负。例如第一个等距样本对各层均值的离差分别为 $-2.75, 4.5, -4.875, 5.75$ 。该样本内六对离差组合中四对的乘积是负数。此外,每个等距样本大都是这种情况。因此,由性质 4, $\rho_{ust} < 0$, 系统抽样的精度高于分层随机抽样。

数据顺序的这种改变不会影响简单随机抽样均值估计的方差 $V(y_{srs})$ 和分层随机抽样均值估计的方差 $V(y_{st})$, 等距抽样均值估计的方差 $V(y_{sy})$ 为:

$$\begin{aligned} V(y_{sy}) &= E(y_{sy} - Y)^2 = \frac{1}{k} \sum_{r=1}^k (y_r - Y)^2 \\ &= \frac{1}{n^2 k} \sum_{r=1}^k (ny_r - nY)^2 = 0.202 \end{aligned}$$

本例中,等距抽样比简单随机抽样和分层随机抽样都更有效。

由上例可见,相对于分层随机抽样和简单随机抽样来说,系统抽样的效率很大

程度上取决于总体性质。即使是相同的总体数据,对于不同的单元排列顺序,就有不同的样本(群)内方差 S_{usy}^2 或相关系数 ρ_{usy} ,从而系统抽样估计量的方差也就不同。因此,要有效地应用系统抽样,必须先了解总体的特征。

§ 7.3 不同特征总体的系统抽样

从上面的讨论中我们知道,系统抽样的精度不仅取决于总体方差的大小,更取决于总体单元的排列顺序。这一节我们进一步研究几种排列特征的总体单元的系统抽样。

一、随机次序排列的总体

在社会经济抽样调查中,许多现象总体的单位是随机排列的,比如居民家计调查中按居民姓氏次序排列的总体单位,农产量调查中按地理区域顺序排队的总体单位,等等。这种按照无关标志排列的总体单元,可以看做是随机排列的。

对于一个有限总体,简单随机抽样的方差是确定的。而系统抽样的方差还取决于单元的排列顺序,对于一个特定的排列,就有一定的数值,因此它是不稳定的,可能大于也可能小于相应的简单随机抽样的方差。比如, N 个总体单元总共有 $N!$ 种不同的排列,从而有 $N!$ 个不同的系统抽样的方差,但可以证明这 $N!$ 个系统抽样方差的平均数恰好等于简单随机抽样的方差。即

$$E(V(y_{sy})) = V(\bar{y}_{sr})$$

因此,平均来说,系统抽样方差与简单随机抽样方差是相等的。在这个意义上,我们说当总体单元按随机顺序排列时,系统抽样的效果等价于简单随机抽样。

当总体单元按随机顺序排列时,就可以采用简单随机抽样的方差作为系统抽样的方差估计:

$$V(y_{sy}) = V(y_{sr}) = \frac{N-n}{Nn} S^2 \quad (7.10)$$

二、线性趋势的总体

(一) 线性趋势的总体

若总体单元按指标值从小到大顺序排列或按某个与其有线性相关的辅助变量的大小顺序排列,此时指标值 Y_i 与单元序号 i 也线性相关,这种按有关标志排列的总体称为线性趋势总体。如图 7.3 所示。

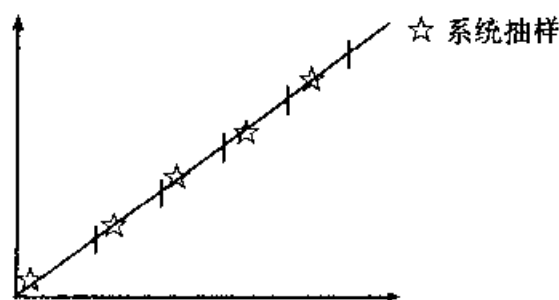


图 7.3 在具有线性趋势的总体中的系统抽样

为了进一步研究这类总体系统抽样的特性,我们先假定一种简单的线性趋势总体,即单元指标 Y_i 值是单元序号 i 的线性函数,即 $Y_i = a + bi (i = 1, 2, \dots, N)$,经过线性变换后,直接假定

$$Y'_i = \frac{Y_i - a}{b} = i, i = 1, 2, \dots, N$$

以下仍用 Y_i 表示 Y'_i 。

下面比较在具有线性趋势总体下,系统抽样的方差 $V(y_{sy})$ 、简单随机抽样的方差 $V(y_{rs})$ 与分层随机抽样的方差 $V(y_{st})$ 。

当 $Y_i = i (i = 1, 2, \dots, N)$ 时,有

$$\begin{aligned} \sum_{i=1}^N Y_i &= \sum_{i=1}^N i = \frac{1}{2}N(N+1) \\ \sum_{i=1}^N Y_i^2 &= \sum_{i=1}^N i^2 = \frac{1}{6}N(N+1)(2N+1) \end{aligned}$$

故总体均值 $\bar{Y} = \frac{1}{2}(N+1)$

$$\text{总体方差 } S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i^2 - N\bar{Y}^2) = \frac{1}{12}N(N+1) \quad (7.11)$$

从而简单随机抽样的方差

$$V(y_{rs}) = \frac{N-n}{nN} S^2 = \frac{1}{12}(k-1)(N+1) \quad (7.12)$$

计算分层随机抽样的方差,由于总体 N 分为 n 层,每层含 k 个单元,且每个单元相差 1,因此线性趋势总体中各层方差 S_w^2 相同,因而

$$S_w^2 = \frac{1}{12}k(k+1)$$

由于考虑的分层随机抽样每层中只抽 1 个样本单元,是按比例分配的,故

$$V(y_s) = \frac{N-n}{Nn} S_u^2 = \frac{1}{12n} (k^2 - 1) \quad (7.13)$$

计算系统抽样的方差,考虑到 k 个可能的系统样本的均值 y_r ,按 r 的不同取值依次都相差 1,因此应用式(7.11)有

$$\frac{1}{k} \sum_{r=1}^k (y_r - Y)^2 = \frac{1}{12} k(k+1)$$

从而

$$V(y_s) = \frac{1}{k} \sum_{r=1}^k (y_r - Y)^2 = \frac{1}{12} (k^2 - 1) \quad (7.14)$$

比较式(7.12)、式(7.13)、式(7.14),可知

$$V(y_{st}) < V(y_s) < V(y_{sr})$$

等号当且仅当 $n = 1$ 时成立。

即系统抽样的方差小于等于简单随机抽样的方差,但大于等于分层随机抽样的方差。因此,一般地,对于线性趋势总体,系统抽样优于简单随机抽样,但比分层随机抽样差

直观来看,总体按线性趋势排列时,这种排列的系统样本内方差增大,故估计量的方差小于简单随机抽样的方差。此外,样本容量为 n 的系统样本可以看做是将总体划分为 n 层,每层抽取一个单位的分层抽样。所不同的是,系统抽样在各层的样本是由第一层中样本单元的位置决定的,如果第一层中样本单元的位置 r 偏低,将导致以后各层样本单元的位置都偏低,样本平均数也偏小;如果第一层中样本单元的位置 r 偏高,将导致以后各层样本单元的位置都偏高,样本平均数也偏大。而分层随机样本的单元在层中的位置是随机的,故由于不同位置对指标值的影响可以抵消一部分,从而使样本平均数的方差进一步减小。

(二) 对线性趋势总体的系统抽样法的改进

虽然以上分析中假设的严格线性趋势排列总体在实际问题中很难成立,但其结论在定性上还是适用的。针对实践中经常出现的线性趋势总体,有必要对系统抽样进行改进,从而提高系统抽样的精度,使系统抽样法有可能达到比分层随机抽样更高的效果。

对线性趋势总体的系统抽样的改进方法主要有两类,一种是抽样方法的改进,如中心位置抽样法、对称系统抽样法等;另一种是估计方法的改进,如首尾校正法。

1. 中心位置抽样法。当总体单元的排列顺序呈线性趋势时,起始单元的位置偏高或偏低会直接影响整个样本的代表性,为提高抽样效率,Madow(1953)建议采用中心位置系统抽样,即初始样本不是随机抽选,而是直接取第一段的 k 个单元中处于中间位置的单元。

当 k 为奇数时, 中点取 $r = \frac{k+1}{2}$;

当 k 为偶数时, 中点取 $r = \frac{k}{2}$ 或 $r = \frac{k}{2} + 1$ 。

这种抽样方法虽然可以提高精度, 但对于一定顺序排列的总体, 样本是确定的, 失去了抽样的随机性。尤其对于连续性调查, 这种抽样会带来不利影响。

例如某学院共有 200 个学生, 要抽 10 个学生做样本, 抽样间距 $k = \frac{N}{n} = \frac{200}{10} = 20$ 。如果采用中心位置抽样法, 起始样本就是第 10 位学生, 其余样本单元依次就是第 30, 50, 70, 90, 110, 130, 150, 170, 190 位学生。

2. 对称系统抽样。对于呈线性趋势排列的总体单元, Sethi 对称系统抽样和 Singn 对称系统抽样都有助于提高系统抽样的精度。

(1) Sethi 对称系统抽样。第一种对称系统抽样方法是由 Sethi(1965) 提出的, 当 $N = nk$, n 为偶数时, 将总体分为 $\frac{n}{2}$ 段, 每段包含 $2k$ 个单元, 在各段内随机选择与两端等距的两个单元作为样本单元, 假设起始随机数为 r ($1 < r \leq k$), 入样的单元为:

$$[r + 2jk, 2(j+1)k - r + 1], j = 0, 1, 2, \dots, \frac{n}{2} - 1$$

当 n 为奇数时, 仍按以上步骤进行, 但到 $j = \frac{n-1}{2} - 1$ 后, 增加靠近终端的一个单元 $[r + (n-1)k]$ 。如图 7.4。

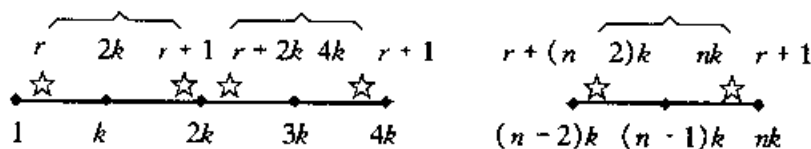


图 7.4 Sethi 对称系统抽样

仍是从 200 位学生中抽取 10 位做样本, 则 $N = 200$, $n = 10$, 抽样间距 $k = \frac{N}{n} = \frac{200}{10} = 20$ 。假设随机抽中 3 为起始单元数, 则样本单元位数依次应该是 3, 38; 43, 78; 83, 118; 123, 158; 163, 198。

(2) Singn 对称系统抽样。Singn(1968) 提出另一种对称系统抽样方法, 当 $N = nk$, n 为偶数时, 假设起始随机数为 r ($1 \leq r \leq k$), 入样的 $\frac{n}{2}$ 对样本单元为:

$$[r + jk, N - r - jk + 1], j = 0, 1, 2, \dots, \frac{n}{2} - 1$$

当 n 为奇数时, 仍按以上步骤进行, 但到 $j = \frac{n-1}{2} + 1$ 后, 增加靠近中间的一个单元 $r + \frac{1}{2}(n-1)k$, 如图 7.5。

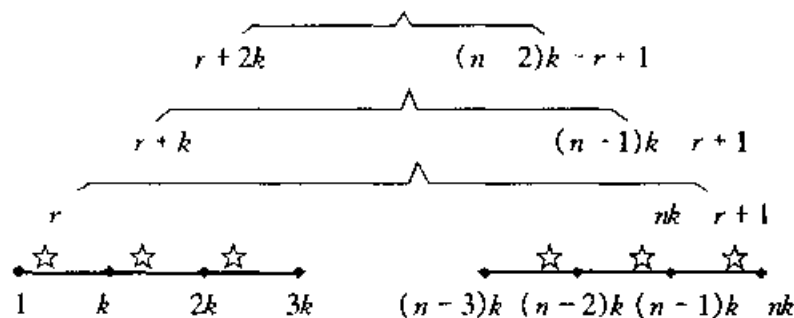


图 7.5 Singn 对称系统抽样

假设从 300 位学生中抽取 15 位做样本, 则 $N = 300, n = 15$, 抽样间距 $k = \frac{N}{n} = \frac{300}{15} = 20$ 。假设随机抽中 3 为起始单元数, 则样本单元位数依次应该是 3, 298; 23, 278; 43, 258; 63, 238; 83, 218; 103, 198; 123, 178; 143。

3. 首尾校正法。首尾校正法通过对首尾两个样本单元赋予不同于其他单元的权数, 从而降低对线性趋势总体的系统抽样的估计偏倚。Yates 首尾校正法主要应用于 $N = nk$ 的情况, Bellhouse 和 Rao 首尾校正法应用于 $N \neq nk$ 的情况。

Yates(1948) 针对 $N = nk$ 的情况, 提出一种用加权平均计算样本均值从而提高精度的方法。这种方法的原理就是对首尾两个样本单元赋予不同于其他单元的权数。假设起始样本单元的编号为 r , 则

$$\text{首样本单元的权数为: } w_1 = \frac{1}{n} + \frac{2r - k - 1}{2(n-1)k} \quad (7.15)$$

$$\text{尾样本单元的权数为: } w_n = \frac{1}{n} - \frac{2r - k - 1}{2(n-1)k} \quad (7.16)$$

$$\text{其他 } n-2 \text{ 个样本单元的权数为: } w_j = \frac{1}{n}, j = 2, \dots, n-1 \quad (7.17)$$

这样, 首尾校正法修正后的总体均值估计量为:

$$y_r = \sum_{j=1}^n w_j y_{rj} \quad (7.18)$$

当总体单元的排列严格线性, 即假定 Y_i 是 i 的线性函数时, 首尾校正法的均值估计量是完全无偏的, 完全不受初始值的影响。

Bellhouse 和 Rao(1975) 将 Yates 的首尾校正法推广到 $N \neq nk$ 的情况。根据 Lahiri 的循环等距抽样(见 7.1 节), 保证 n 为常数。然后按照总体单元原有顺序确

定首尾单元,对其赋予不同于其他单元的权数。

如果初始单元编号 r 较小,满足 $r + (n - 1)k \leq N$,则所有 n 个样本单元都不经过单元 N ,相应的权数如下:

$$\text{首样本单元的权数为: } w_1 = \frac{1}{n} + \frac{2r + (n - 1)k - (N + 1)}{2(n - 1)k} \quad (7.19)$$

$$\text{尾样本单元的权数为: } w_n = \frac{1}{n} - \frac{2r + (n - 1)k - (N + 1)}{2(n - 1)k} \quad (7.20)$$

$$\text{其他 } n - 2 \text{ 个样本单元的权数为: } w_j = \frac{1}{n}, j = 2, \dots, n - 1 \quad (7.21)$$

如果初始单元编号 r 较大,满足 $r + (n - 1)k > N$,则有样本单元越过单元 N ,假设越过单元 N 的样本单元有 n_2 个,相应的权数如下:

$$\text{首样本单元的权数为: } w_1 = \frac{1}{n} + \frac{2r + (n - 1)k - (N + 1) - 2n_2 \frac{N}{n}}{2(N - k)} \quad (7.22)$$

$$\text{尾样本单元的权数为: } w_n = \frac{1}{n} - \frac{2r + (n - 1)k - (N + 1) - 2n_2 \frac{N}{n}}{2(N - k)} \quad (7.23)$$

$$\text{其他 } n - 2 \text{ 个样本单元的权数为: } w_j = \frac{1}{n}, j = 2, \dots, n - 1 \quad (7.24)$$

【例 7.4】 总体有 23 个单位,拟抽取 $n = 5$,则 $k = \frac{N}{n} = 4.6$,取与之最近的整数 $k = 5$ 。然后在总体中随机抽取一个单位作为起点,假设抽中 $r = 19$,样本单位的顺序编号分别为:19,1,6,11,16。首样本单元为 y_1 ,尾单元为 y_{19} 。求相应单元的权数。

解:由于 $n_2 = 4, N = 23, n = 5, k = 5, r = 19$

$$\begin{aligned} \text{首样本单元 } y_1 \text{ 的权数为: } w_1 &= \frac{1}{n} + \frac{2r + (n - 1)k - (N + 1) - 2n_2 \frac{N}{n}}{2(N - k)} \\ &= 0.1222 \end{aligned}$$

$$\begin{aligned} \text{尾样本单元 } y_{19} \text{ 的权数为: } w_n &= \frac{1}{n} - \frac{2r + (n - 1)k - (N + 1) - 2n_2 \frac{N}{n}}{2(N - k)} \\ &= 0.2778 \end{aligned}$$

其他 3 个样本单元的权数为:0.2

三、周期波动的总体

周期性波动是指总体单元指标值按其顺序呈周期性变化。例如商店的日销售额以 7 天为周期变化,一般周末为销售高峰期,周一、周二下降;城市交通量以 24 小时为周期变化,上下班时间为高峰期。典型的周期性波动如图 7.6 所示。

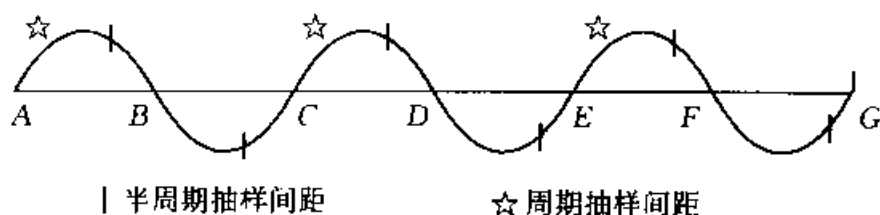


图 7.6 周期波动总体系统抽样示意图

对于周期性波动总体,使用系统抽样一定要特别注意。系统抽样的估计效果与抽样间距 k 及单元指标值的变化周期直接有关。

如图 7.6,如果抽样间距等于周期(AC) 倍数,任意系统样本内的单位都会有相同数值,此时系统样本的代表性最差,仅相当于从总体中随机抽取了一个样本;而且不同系统样本间的差异很大,会导致很大的估计方差。

如果抽样间距等于半周期(AB) 倍数,系统样本内的单位会依次高于、低于中线,系统抽样会得到无偏的均值估计,估计方差也会大大减少。

对于周期倍数与半周期倍数之外的系统抽样间距,抽样的效果主要取决于抽样间距与周期长度的关系。现实中,对于含有周期影响的总体,如果已经掌握其周期结构,合理选择系统抽样间距 k ,使样本中包含周期中许多有代表性的指标值,可以大大缩小估计量的方差,系统抽样的效果会相当好。但如果对总体的周期结构不甚了解,简单随机抽样和分层随机抽样的效果可能会更好。

§ 7.4 系统抽样的方差估计

系统抽样法的缺点之一,就是很难得出估计方差的无偏估计。本节介绍几种形式相对简单的方差估计方法,这些方差估计方法只能进行近似估计,而且不同的方法适用于不同的总体模型。

一、等概系统抽样的方差估计

为方便起见,将系统样本观测值按其在总体中的顺序记为 y_1, y_2, \dots, y_n , 我们

讨论用 $y_{sy} = \frac{1}{n} \sum_{i=1}^n y_i$ 估计总体均值 \bar{Y} 时的方差 $V(\bar{y}_{sy})$ 的估计。

(一) 系统样本来自随机排列总体

假设系统样本来自随机排列总体,系统样本可近似视为简单随机样本,从而可以采用简单随机抽样下抽样方差的无偏估计:

$$v_1 = \frac{1-f}{n} f_s^2 = \frac{N-n}{nN} \frac{1}{n-1} \sum_{i=1}^n (y_i - y_{sy})^2 \quad (7.25)$$

(二) 系统样本分层随机抽取

如果把系统样本看成从各层抽取两个单位的分层随机抽样,可采用以下方法。

1. 从第二个样本单元开始,每个样本单元与前一个样本单元组成一对,共 $n-1$ 对,第 i 对样本单元的方差估计为 $\frac{1}{2}(y_{i+1} - y_i)^2$,因此对 $n-1$ 个 $\frac{1}{2}(y_{i+1} - y_i)^2$ 进行平均,再乘以 $\frac{1-f}{n}$,得 $V(\bar{y}_{sy})$ 的估计:

$$v_2 = \frac{1-f}{n} \times \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 = \frac{N-n}{2n(n-1)N} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 \quad (7.26)$$

2. 设 n 为偶数,将样本单元按顺序两两分成一组,共 $\frac{n}{2}$ 组,第 i 对样本单元的方差估计为 $\frac{1}{2}(y_{2i} - y_{2i-1})^2$,将这 $\frac{n}{2}$ 个方差估计值进行平均,再乘以 $\frac{1-f}{n}$,从而得到

$$v_3 = \frac{1-f}{n} \times \frac{2}{n} \times \frac{1}{2} \sum_{i=1}^{\frac{n}{2}} (y_{2i} - y_{2i-1})^2 = \frac{N-n}{n^2 N} \sum_{i=1}^{\frac{n}{2}} (y_{2i} - y_{2i-1})^2 \quad (7.27)$$

(三) 系统样本来自线性趋势总体

假设系统样本来自线性趋势总体,即 $Y_i = a + b_i + e_i (i = 1, 2, \dots, N)$, $E(e_i) = 0, F(e_i^2) = \sigma^2, E(e_i e_j) = 0$,进行 Yates 首尾校正法后

$$\hat{Y} = y + \frac{2r-k-1}{2(n-1)k} (y_r - y_r + (n-1)k) \quad (7.28)$$

其抽样方差无偏估计为:

$$v_4 = \frac{k-1}{nk} \left[\frac{1}{n} - \frac{(2r-k-1)^2}{2(n-1)^2 k^2} \right] \sum_{i=1}^{\frac{n-2}{2}} \frac{(y_i - 2y_{i+1} + y_{i+2})^2}{6(n-2)} \quad (7.29)$$

当 n 较大时,中括号项可忽略。但当线性模型存在异方差时, $v(\hat{Y})$ 不再是无偏估计。

(四) 样本量为 n 的系统样本分成 m 个子样本独立抽取

样本量为 n 的系统样本分成 m 个子样本独立抽取, 每个子样本仍用系统抽样法, 样本量为 $n' = \frac{n}{m}$, 抽样间距为 $k' = mk$, 每个子样本的起始值独立抽取。记第 α 个子样本的均值为 y_α , 总体均值的估计值为:

$$\hat{Y} = \frac{1}{m} \sum_{\alpha=1}^m y_\alpha \quad (7.30)$$

则 $V(y_\alpha)$ 的估计是:

$$v_5 = \frac{1}{m(m-1)} \sum_{\alpha=1}^m (y_\alpha - \hat{Y})^2 \quad (7.31)$$

$m=2$ 时, 上式可简化为 $\frac{(\bar{y}_1 - \bar{y}_2)^2}{4}$ 。

上述 m 个子样本的抽取是相互独立的, 样本单元也有可能重复, 所以可以采取将样本量为 n ($n = mn$), 抽样间隔为 $k \left(\frac{k'}{m} \right)$ 的系统样本分成 m 个系统子样本, 每个子样本的样本量为 n' , 间隔为 k' 。但这样的子样本相互不独立, v_5 也不再是无偏的了。这种方法称为交叉子样本法, 也称随机组法。

以上估计方法大都是建立在一定的假设模型之上的, 不同的模型反映不同特征的总体。因此, 在实践中只有所研究的总体符合假设模型时, 才可以用相应的抽样方差公式来估算系统抽样方差。

一般情况下, 对于随机排列总体, 以上估计方法的效果都不错, 但简单随机抽样的方差估计 v_1 最简单, 故为最佳选择。对于线性趋势总体, v_2 和 v_3 的效果最好, v_2 相对更适用于样本量较小的情况; 对于周期波动总体, 上述估计量都不太理想。当抽样间距为周期倍数时, 这些估计量都偏小; 而当抽样间距为半周期奇数倍时, 这些估计量又都偏大。

如果对总体背景不甚了解, 建议采用 v_2 和 v_3 。这两个估计量普遍适用于随机排列、线性趋势和周期波动总体, 而且效果也不错。

基于交叉子样本的方差估计虽然也适用于各种类型的总体, 但实际操作并不方便, 而且对于线性趋势总体和自相关总体效果也不是很好。此外, 使用交叉子样本法时, 抽取的系统样本个数不能太多, 所以会带来经济效率损失。

二、不等概系统抽样的方差估计

以上几节介绍的系统抽样法都是等概系统抽样, 每个单元の入样概率是相等的。但在实际应用中, 不等概系统抽样的应用也很广泛。不等概系统抽样结合了系

统抽样方便易行和不等概抽样的高效率,是不放回不等概抽样方法中最受欢迎的方法之一。

不等概系统抽样中每个单元的入样概率不相等。对于按一定顺序排列的 N 个总体单元,假设 π_i ($i = 1, 2, \dots, N$) 是一组入样概率,且 $\sum_{i=1}^N \pi_i = n$ 。不等概系统抽样的一般方法就是先在 $[0, 1]$ 区间内随机抽取一随机数 r ,则满足以下条件的总体中的第 $i_0, i_1, i_2, \dots, i_{n-1}$ 个单元即为抽中的样本单元。

$$\sum_{j=1}^k \pi_{i_j} < r + k, \sum_{j=1}^{k+1} \pi_{i_j} \geq r + k, k = 0, 1, 2, \dots, n-1$$

当 $\pi_i \leq 1$ 时,抽样是严格不放回的。

(一) 估计量及其方差

对于不等概系统抽样,对总体总和 Y 的估计可采用通常不放回的不等概抽样中的 Horvitz - Thompson 估计量:

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} \quad (7.32)$$

对于 π PS 系统抽样,有

$$\hat{Y}_{HT} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i} \quad (7.33)$$

\hat{Y}_{HT} 是无偏的,其方差可表达为:

$$V(\hat{Y}_{HT}) = \sum_{i=1}^N \frac{1}{\pi_i} \pi_i Y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j \quad (7.34)$$

当 n 固定时,

$$V(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j>i}^n (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 \quad (7.35)$$

(二) 不等概系统抽样的方差估计

不等概系统抽样对总体总和 Y 的估计可采用不放回的不等概抽样中的 Horvitz - Thompson 估计量 \hat{Y}_{HT} ,但由于 π_{ij} 的计算极为复杂,且有可能为零,其方差估计式显然并不适于系统样本。下面我们介绍几种不等概系统抽样的方差估计方法。

1. 是将不放回的 π PS 系统样本作为放回的 PPS 样本处理,可得到以下的方差估计形式:

$$v_6 = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{Y}_{HT} \right)^2 = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{ny_i}{\pi_i} - \hat{Y}_{HT} \right)^2 \quad (7.36)$$

2. 因为实际抽样是不放回的,为此应考虑乘以有限总体修正系数 $1 - f$ 。由于这里的单元实际上是不平等的,因此 f 不是简单地等于 $\frac{n}{N}$ 。我们使用 f 的以下估计:

$$\hat{f} = \frac{1}{n} \sum_{i=1}^n \pi_i$$

因而得到方差估计量的一种方式:

$$v_7 = (1 - \hat{f}) v_6 = \frac{1}{n(n-1)} \sum_{i=1}^n \frac{\pi_i}{n} \sum_{j=1}^n \left(\frac{ny_i}{\pi_i} - \hat{Y}_{HT} \right)^2 \quad (7.37)$$

3. 用相邻样本单元差值的平方和来表示方差,这里用 $\frac{ny_i}{\pi_i}$ 代替等概率情形的 y_i ,得到

$$v_8 = \frac{1 - \hat{f}}{2n(n-1)} \sum_{i=2}^n \left(\frac{ny_i}{\pi_i} - \frac{ny_{i-1}}{\pi_{i-1}} \right)^2 \quad (7.38)$$

$$v_9 = \frac{1 - \hat{f}}{n^2} \sum_{i=1}^{\frac{n}{2}} \left(\frac{ny_{2i}}{\pi_{2i}} - \frac{ny_{2i-1}}{\pi_{2i-1}} \right)^2 \quad (7.39)$$

4. 将样本量为 n 的系统样本随机分成 m 个子样本,每个子样本样本量为 n'

$\frac{n}{m}$, 记第 α 个子样本对总和的 HT 估计为:

$$\hat{Y}_\alpha = \frac{m}{n} \sum_{i=1}^{\frac{n}{m}} \frac{ny_{\alpha i}}{\pi_i} \quad (7.40)$$

则不等概系统抽样方差的估计是

$$v_{10} = \frac{1}{m(m-1)} \sum_{\alpha=1}^m (\hat{Y}_\alpha - \hat{Y}_{HT})^2 \quad (7.41)$$

同样,以上估计方法适用于不同特征的总体。

对于随机排列总体,以上估计方法的效果都不错, v_7 为较好选择。对于线性趋势总体, v_8 和 v_9 的效果最好, v_8 相对更适用于样本量较小的情况。与等概系统抽样相似, v_{10} 的效果不太理想,一般不推荐使用。

小 结

本章介绍了实践中最常用的系统抽样方法。系统抽样既可以看成是一种特殊

的整群抽样,又可以看成是一种特殊的分层抽样,它的最大优点是简便易行。此外,在了解总体特征的前提下,有效地应用系统抽样还可以得到很高的精度。反之,如果缺乏对总体的认识,比如直接对隐藏有周期性波动的总体进行等距抽样,得到的系统样本的代表性可能会很差。

一般地,对于线性趋势总体,系统抽样优于简单随机抽样,但比分层随机抽样差。针对实践中经常出现的线性趋势总体,有必要对系统抽样进行改进,改进后的系统抽样有可能达到比分层随机抽样更好的效果。

系统抽样的方差估计较为复杂,一般系统抽样难以找到设计意义下的无偏估计量。系统抽样方差的近似估计方法很多,但这些方法都有各自适用的总体模型。在实践中无论是选择系统抽样方法,还是确定系统抽样方差的估计方法,都有必要先了解所研究总体的特征。

本章附录 不同特征总体系统抽样的性质证明

1. 证明性质 2: 用样本(群)内方差 S_{usy}^2 表示系统抽样估计量的方差:

$$V(Y_{sy}) = \frac{(N-1)}{N} S^2 - \frac{k(n-1)}{N} S_{usy}^2 \quad (7.5)$$

式中, $S^2 = \frac{1}{N-1} \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - Y)^2$ 为总体方差; $S_{usy}^2 = \frac{1}{k(n-1)} \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - \bar{y}_r)^2$ 为样本(群)内方差。

证明: 将总体平方和按照全部可能的系统样本(表 7.2 中的行)进行分解, 得到

$$\begin{aligned} (N-1)S^2 &= \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - Y)^2 \\ &= n \sum_{r=1}^k (y_r - \bar{Y})^2 + \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - \bar{y}_r)^2 \\ &= \frac{nk}{k} \sum_{r=1}^k (y_r - Y)^2 + \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - y_r)^2 \end{aligned}$$

根据 $V(y_{sy})$ 定义, 且 $nk = N$, 得

$$\begin{aligned} V(y_{sy}) &= \frac{(N-1)}{N} S^2 - \frac{1}{N} \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - y_r)^2 \\ &= \frac{(N-1)}{N} S^2 - \frac{k(n-1)}{N} S_{usy}^2 \end{aligned}$$

式中, $S_{usy}^2 = \frac{1}{k(n-1)} \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - y_r)^2$, 为样本(群)内方差。

2. 证明性质 3: 用样本(群)内相关系数 ρ_{usy} 表示系统抽样估计量的方差:

$$V(y_{sy}) = \frac{S^2}{n} \left(\frac{N-1}{N} \right) [1 + (n-1)\rho_{usy}] \quad (7.6)$$

式中, ρ_{usy} 为样本(群)内相关系数。

$$\rho_{usy} = \frac{E(y_{rj} - Y)(y_{ru} - Y)}{E(y_{rj} - Y)^2} = \frac{2}{(n-1)(N-1)S^2} \sum_{r=1}^k \sum_{j < u}^n (y_{rj} - Y)(y_{ru} - Y) \quad (7.7)$$

证明: 由于系统抽样可以看做是一种特殊的整群抽样, 而且群的大小相等。因此, 可以直接利用整群抽样的公式表示。由(6.11)式, 整群抽样总体均值的估计量 \bar{y} 的方差可表示为:

$$V(\bar{y}) = \frac{1-f}{n} \cdot \frac{NM-1}{M^2(N-1)} \cdot S^2 \cdot [1 + (M-1)\rho]$$

系统抽样与整群抽样的参数对照见下表:

	总体单元数	群内单元数	总体群数	样本群数	总体均值估计量	群内相关系数
系统抽样	N	n	k	1	\bar{y}_{sy}	ρ_{usy}
整群抽样	NM	M	N	n	\bar{y}	ρ

因此有 $V(y_{sy}) = \frac{S^2}{n} \left(\frac{N-1}{N} \right) [1 + (n-1)\rho_{usy}]$

3. 证明性质 4: 系统抽样估计量的方差可以用层内方差 S_{ust}^2 和 ρ_{ust} 表示:

$$V(y_{sy}) = \frac{S_{ust}^2}{n} \left(\frac{N-1}{N} \right) [1 + (n-1)\rho_{ust}] \quad (7.8)$$

式中, S_{ust}^2 为层内方差; $S_{ust}^2 = \frac{1}{n(k-1)} \sum_{j=1}^n \sum_{r=1}^k (y_{rj} - y_{\cdot j})^2$; ρ_{ust} 为同一系统样本内对层均值离差的相关系数。

$$\rho_{ust} = \frac{E(y_{rj} - y_{\cdot j})(y_{ru} - y_{\cdot u})}{E(y_{rj} - y_{\cdot j})^2} = \frac{2}{n(n-1)(k-1)S_{ust}^2} \sum_{r=1}^k \sum_{j < u}^n (y_{rj} - y_{\cdot j})(y_{ru} - y_{\cdot u}) \quad (7.9)$$

证明: 由于系统抽样可看做是一种特殊的分层抽样(见表 7.2), 从每层的固定位置抽取一个单元的分层抽样, 而且各层大小相等。记第 j 层的均值为:

$$y_{\cdot j} = \frac{1}{k} \sum_{r=1}^k y_{rj}, j = 1, 2, \dots, n$$

系统样本对层均值离差的相关系数:

$$\rho_{ust} = \frac{E(y_{rj} - y_{\cdot j})(y_{ru} - y_{\cdot u})}{E(y_{rj} - y_{\cdot j})^2}$$

$$= \frac{2}{n(n-1)(k-1)S_{ust}^2} \sum_{r=1}^k \sum_{j < u}^n (y_{rj} - y_{\cdot j})(y_{ru} - y_{\cdot u})$$

根据(7.4)式,

$$V(y_{sy}) = \frac{1}{k} \sum_{r=1}^k (y_r - \bar{Y})^2$$

两边同乘以 n^2k , 有

$$\begin{aligned} n^2kV(y_{sy}) &= n^2 \sum_{r=1}^k (y_r - \bar{Y})^2 \\ &= \sum_{r=1}^k \left[\sum_{j=1}^n (y_{rj} - y_{\cdot j}) \right]^2 \\ &= \sum_{j=1}^n \sum_{r=1}^k (y_{rj} - \bar{y}_{\cdot j})^2 + 2 \sum_{r=1}^k \sum_{j < u}^n (y_{rj} - y_{\cdot j})(y_{ru} - y_{\cdot u}) \\ &= n(k-1)S_{ust}^2 + n(n-1)(k-1)S_{ust}^2\rho_{ust} \\ &\quad (N-n)S_{ust}^2 + (n-1)(N-n)S_{ust}^2\rho_{ust} \end{aligned}$$

因而

$$V(y_{sy}) = \frac{S_{ust}^2}{n} \left(\frac{N-n}{N} \right) [1 + (n-1)\rho_{ust}]$$

习 题

1. 系统抽样设计的原理是什么? 系统抽样与整群抽样、分层抽样的关系如何?
2. 假定系统样本的平均数为 y_{sy} , 试证明:

$$(1) V(y_{sy}) = \frac{(N-1)}{N} S^2 - \frac{k(n-1)}{N} S_{usy}^2$$

式中, S^2 为总体方差; $S_{usy}^2 = \frac{1}{k(n-1)} \sum_{r=1}^k \sum_{j=1}^n (y_{rj} - \bar{y}_r)^2$ 为样本内方差。

(2) 在相同样本量的情况下, 当且仅当 $S_{usy}^2 > S^2$ 时, 系统抽样法优于简单随机抽样。

3. 回答下列问题:

(1) 某班级共 40 人, 若样本量 $n = 7$, 随机起点 $r = 5$, 请用循环等距抽样方法列出样本单元序号。

(2) 某班级共 35 人,若样本量 $n = 7$,随机起点 $r = 5$,请用 Sethi 对称系统抽样和 Singn 对称系统抽样列出样本序号

4. 某地的 360 户(编号 1 ~ 360)的总体,在档案中按户主的姓氏字母次序排列 下列号码是户主为汉族的住户的号码:

28,31	33,36	41,44,45,47,55,56,58,68,69,82,83,85,86,89	94,
98,99,101,107	110,114,154,156,178,223,224,296,298	300,	
302	304,306	323,325	331,333,335 339,341,342

为了估计户主为汉族的住户在全部住户中所占的比重,每 8 户抽 1 户,取得一个等距样本。试将这一等距样本的精确度与同样含量的简单随机样本的精确度加以比较。

5. 有一个紧邻地区,其居民分别是汉族、回族和蒙古族。还有一本最近的居民册,册内每一户的人是依下列顺序登记的:丈夫、妻子、孩子(按年龄排列)、其他人,各户是沿街道按顺序排列的,每户平均有 5 口人。有两种抽样方案:

(1) 在户口册中每 5 人抽 1 人,可以得到一个系统样本;

(2) 按 20% 的比例抽取一个简单随机样本。

现在要从这两种样本中选择一种样本。在下述的三种指标中,你认为哪一种指标采用等距样本有希望取得更好的精确度呢?并请说明理由。(1) 汉族所占的比例;(2) 男性所占的比例;(3) 孩子所占的比例。

6. 在一条街上 13 户的户口册中,将所有的居民列成下表(M 为男性成人;F 为女性成人;m 为男孩;f 为女孩):

为了估计下列各项指标:(1) 男性所占比例;(2) 孩子所占比例;(3) 具有某种职业的住户中人员的比例(1,2,3,12,13 这几户是职业性住户)。现从每 5 个人中抽 1 人得到一个系统样本,又按 20% 的比例抽取一个简单随机样本,请比较这两种样本的方差。这些结果是否说明习题 5 中你的回答是正确的?系统样本的排列方法是每户从上到下依次排列。

1	2	3	4	5	6	7	8	9	10	11	12	13
M	M	M	M	M	M	M	M	M	M	M	M	M
F	F	F	F	F	F	F	F	F	F	F	F	F
f	f	m		m	f	f	m	m	m	f	f	
m	m	f		m	m	f	f		f	m		
f	f			f			m					

7. 假设总体, 相应指标值排列顺序为 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15。

(1) 考虑 $n = 3$ 的直线系统抽样, 计算系统抽样的实际方差, 与样本量相同的简单随机抽样进行比较

(2) 若要求抽样间距 $k = 4$, 样本均值是否为总体均值的无偏估计? 它在何时有偏? 何时无偏?

8. 检查某书稿上的错字, 每 5 页检查一页上的错字数, 系统抽取 30 页样品后的检查结果如下:

10	8	6	5	9	8	8	5	9	9
9	10	4	3	1	2	3	4	0	6
3	5	0	3	0	0	4	0	8	0

(1) 试估计这本书稿的平均错字数;

(2) 用合并层方法估计抽样方差;

(3) 用连续差方法估计抽样方差;

(4) 用交叉子样本法估计抽样方差。



第 8 章

多阶段抽样

整群抽样具有样本比较集中的特点,因此它可以节省调查费用,而且便于组织实施,回答率通常也较高。但是由于群内单元通常具有相似性(表现为群内相关系数大于零),尤其是当群比较大时,人们自然会想到没有必要对群内所有单元都进行调查,而是对群内单元进行再抽样,对部分被抽中的单元进行调查,这就是实际工作中常用的多阶段抽样。

本章共分四节,第一节将介绍多阶段抽样的定义、作用以及推算原理,第二节介绍初级单元大小相等时两阶段抽样估计量及其性质,第三节介绍初级单元大小不等时两阶段抽样估计量及其性质,第四节介绍样本量的确定以及多阶段抽样的问题。

§ 8.1 引言

在一项某市居民对香皂颜色喜好的调查中,设计者打算采用入户调查的方式。为节省差旅费,希望样本能够比较集中,因此准备采用整群抽样。方案设计者手头有一份全市各行政区所属的街道名单。如果采用整群抽样,以街道作为抽样单元

(群) 进行抽样, 并调查样本街道所有的居民户, 则群内调查的工作量太大; 如果以居委会作为抽样单元(群), 则群内调查的工作量会小得多, 但以居委会作为抽样单元, 需要事先掌握各街道下属居委会的名单。

由于时间和经费的限制, 编制全市的居委会名单已经来不及了, 设计者考虑对上面的方案进行一些改变。首先, 他决定只抽取部分街道并建立其所属的居委会名单, 并抽出部分居委会; 其次, 他觉得对样本居委会中的每户家庭都进行调查不仅费时而且没有必要, 因此决定只调查其中的部分居民户。

经过修改后的方案是, 首先对街道进行抽样, 在被抽中的街道中分别建立所属的居委会名单并分别抽出部分居委会, 在被抽中的居委会中抽取部分居民户作为样本并进行调查。这个方案的抽样是分三个阶段进行的, 即先抽出样本街道, 再从中抽出样本居委会, 最后从样本居委会中抽出样本居民户。这时的抽样就运用了多阶段抽样的方法。

一、定义与作用

(一) 多阶段抽样的定义

先在总体个单元(初级单元)中抽出个样本单元, 并不对这个样本单元中的所有下一级单元(二级单元)都进行调查, 而是在其中再抽出若干个二级单元并进行调查。这种抽样方法称为二阶段抽样。同样的道理, 还可以有三阶段抽样、四阶段抽样等。对于二阶段以上的抽样, 称为多阶段抽样。

例如, 以全国为总体进行某项调查, 可以定义全国的县为初级单元, 乡镇为二级单元, 自然村为三级单元, 户为四级单元等。在全国抽取若干样本县, 对样本县再抽若干样本乡镇, 在样本乡镇中, 抽取若干自然村, 在自然村中抽取样本户, 这是一个四阶段抽样问题。又如, 关于某市居民对香皂颜色喜好的调查, 采用的是三阶段抽样。

在实际使用多阶段抽样时, 各阶段的定义可以根据行政管理级别确定, 如上面的街道、居委会、居民户。但并不是所有调查都按这种方式进行, 如从城市抽街道就跳过了区级行政机构, 还可以跳过居委会直接抽居民户等。具体工作中如何决定各阶段的抽样单元, 要根据抽样组织管理的方便和实际的可能进行。

(二) 多阶段抽样的优点

在大范围的抽样调查中, 多阶段抽样是一种常用的抽样技术。我们已经讨论了整群抽样, 整群抽样的主要优点是样本比较集中、便于调查、节省经费等, 但由于群内单元的相似性, 使得整群抽样的抽样方差通常比相同样本量的简单随机抽样的抽样方差大。另外, 在群比较大的时候, 如果对群内每个单元都进行调查, 则体现不

出抽样调查的优点。因此,人们很自然地想到,可以对样本群中的下一级单元再进行抽样,这就提出了多阶段抽样的问题。

多阶段抽样一方面保持了整群抽样的样本比较集中、便于调查、节省费用等优点,同时又避免了对小单元过多调查造成的浪费,充分发挥抽样调查的优点。

多阶段抽样的另一个优点是不需要编制所有小单元的抽样框。抽取初级单元时,只需编制初级单元的抽样框,对被抽中的初级单元,再去编制二级单元抽样框,依此类推,每阶段只需编制该阶段的抽样框,从而大大降低编制抽样框的工作量。对于有些调查问题,抽样框的变动非常频繁,待抽样框整理完毕后,可能与实际情况相去甚远,这时,多阶段抽样是解决这类问题的一个办法。全国范围内的调查一般都用多阶段调查的技术,即使是在某个城市范围内的居民调查,对一家调查公司而言,不可能也没有必要编制全市的居民名单抽样框,多阶段抽样方法就可以解决这一问题。

二、抽选方法与推断原理

多阶段抽样每一个阶段的抽样可以相同,也可以不同,它通常与分层抽样、整群抽样、系统抽样结合使用。一般来说,当初级单元大小相同时,第一阶段的抽样采用简单随机抽样;当初级单元大小不同时,第一阶段的抽样采用不等概抽样。

如果两阶段抽样中所有初级单元都被抽中,在每个初级单元中抽取部分二级单元,则这时的抽样就成为分层抽样。如果对初级单元进行抽样,并且样本初级单元中的所有二级单元都被抽中,则这时的抽样就成为整群抽样。

实际工作中,多阶段抽样通常和整群抽样结合使用,即前几阶是多阶段抽样,最后一阶为整群抽样。例如,关于居民对香皂颜色喜好的调查,前两阶抽街道、居委会,最后一阶抽居民户,并对样本居民户中的所有居民都进行调查,这时的居民户就是由其所属的居民组成的一个群。

多阶段抽样时,抽样是分步进行的,因此,讨论估计量 $\hat{\theta}$ 的均值及其方差时需要分阶段进行,这要用到下面的性质1。

性质1 对于两阶段抽样,有

$$E(\hat{\theta}) = E_1 E_2(\hat{\theta}) \quad (8.1)$$

$$V(\hat{\theta}) = V_1[E_2(\hat{\theta})] + E_1[V_2(\hat{\theta})] \quad (8.2)$$

式中, E_2, V_2 为在固定初级单元时对第二阶抽样求均值和方差; E_1, V_1 为对第一阶抽样求均值和方差。

性质 1 可以推广到多阶段抽样的情形,例如对于三阶段抽样,有

$$E(\hat{\theta}) = E_1 E_2 E_3(\hat{\theta}) \quad (8.3)$$

$$V(\hat{\theta}) = V_1[E_2 E_3(\hat{\theta})] + E_1[V_2[E_3(\hat{\theta})] + E_1 E_2[V_3(\hat{\theta})]] \quad (8.4)$$

§ 8.2 初级单元大小相等的二阶抽样

首先考虑初级单元中二级单元规模相等的情形。对于初级单元大小不等的情形,可以通过分层,将大小近似的初级单元分到一层,则层内的二阶抽样就可以按初级单元大小相等的方式来处理。

第一阶段在总体 N 个初级单元中,以简单随机抽样抽取 n 个初级单元,第二阶段在被抽中的初级单元包含的 M 个二级单元中,以简单随机抽样抽取 m 个二级单元,即最终接受调查的单元。

例如,某个新开发的小区拥有相同户型的 15 个单元的楼盘,居民已经陆续搬入新居,每个单元住有 12 户居民,为调查居民家庭装潢情况,准备从 180 户居民户中抽取 20 户进行调查。如表 8.1。

表 8.1 二阶段抽样示意表

编号	单元	房 号											
1	一栋 A 座	1	2*	3*	4*	5	6	7	8	9	10*	11	12
2	一栋 B 座	1	2	3	4	5	6	7	8	9	10	11	12
3	一栋 C 座	1	2	3	4	5	6	7	8	9	10	11	12
4	一栋 A 座	1	2	3	4	5	6	7	8	9	10	11	12
5	一栋 B 座	1	2	3	4	5	6	7	8	9	10	11	12
6	一栋 C 座	1*	2	3	4	5	6*	7	8	9*	10	11*	12
7	一栋 A 座	1	2	3	4	5	6	7	8	9	10	11	12
8	一栋 B 座	1	2	3	4	5	6	7	8	9	10	11	12
9	一栋 C 座	1	2	3	4	5*	6	7*	8*	9	10*	11	12
10	四栋 A 座	1	2	3	4	5	6	7	8	9	10	11	12
11	四栋 B 座	1	2	3	4	5	6	7	8	9	10	11	12
12	四栋 C 座	1	2	3	4	5*	6	7*	8*	9	10	11*	12
13	五栋 A 座	1	2	3	4*	5	6*	7*	8	9	10	11*	12
14	五栋 B 座	1	2	3	4	5	6	7	8	9	10	11	12
15	五栋 C 座	1	2	3	4	5	6	7	8	9	10	11	12

* 为被抽中的房号

我们可以利用二阶抽样方法。这时,初级单元有 15 个,每个初级单元拥有的二级单元为 12 个。首先将单元从 1 到 15 编号,在 15 个单元中随机抽取部分单元,抽取了 5 个单元,分别是 1,6,9,12,13 号;然后在被抽中的单元中,分别独立地随机抽取若干户居民并进行调查,即在这 5 个单元中,分别在 12 户居民户中随机抽取 4 户

一、符号说明

初级单元和初级单元拥有的二级单元个数: N, M

第一阶段和第二阶段抽样的样本量: n, m

第 i 个初级单元中的第 j 个二级单元的观测值: Y_{ij}

样本中第 i 个初级单元中的第 j 个二级单元的观测值: y_{ij}

第一阶段和第二阶段的抽样比: $f_1 = \frac{n}{N}, f_2 = \frac{m}{M}$

第 i 个初级单元按二级单元的平均值: $Y_i = \frac{1}{M} \sum_{j=1}^M Y_{ij}, y_i = \frac{1}{m} \sum_{j=1}^m y_{ij}$

按二级单元的平均值: $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$

初级单元间的方差: $S_1^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2, s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2$

初级单元内的方差: $S_2^2 = \frac{1}{N(M-1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - Y_i)^2$

$$s_2^2 = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - y_i)^2$$

由 S_2^2 的表达式注意到,若记

$$S_{2i}^2 = \frac{1}{M-1} \sum_{j=1}^M (Y_{ij} - Y_i)^2 \quad (8.5)$$

则有

$$S_2^2 = \frac{1}{N} \sum_{i=1}^N S_{2i}^2 \quad (8.6)$$

即 S_2^2 是 S_{2i}^2 的平均值。

同理,若记

$$s_{2i}^2 = \frac{1}{m-1} \sum_{j=1}^m (y_{ij} - y_i)^2 \quad (8.7)$$

则有

$$s_2^2 = \frac{1}{n} \sum_{j=1}^n s_{2j}^2 \quad (8.8)$$

二、估计量及其性质

(一) 总体均值的估计

性质 2 对于初级单元大小相等的二阶抽样, 如果两个阶段都是简单随机抽样, 且对每个初级单元, 第二阶抽样是相互独立进行的, 则对总体均值 \bar{Y} 的无偏估计为:

$$\hat{\bar{Y}} = \bar{\bar{y}} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m y_{ij} \quad (8.9)$$

其方差为:

$$V(\bar{y}) = \frac{1}{n} \frac{f_1}{S_1^2} + \frac{1-f_2}{nm} S_2^2 \quad (8.10)$$

$V(\bar{y})$ 的无偏估计为:

$$v(\bar{y}) = \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{nm} s_2^2 \quad (8.11)$$

【例 8.1】 欲调查 4 月份 100 家企业的某项指标, 首先从 100 家企业中抽取了一个含有 5 家样本企业的简单随机样本, 由于填报一个月的数据需要每天填写流水账, 为了减轻样本企业的负担, 调查人员对这 5 家企业分别在调查月内随机抽取 3 天作为调查日, 要求样本企业只填写这 3 天的流水账。调查的结果如表 8.2。

表 8.2 对 5 家企业的调查结果

样本企业	第一日	第二日	第三日
1	57	59	64
2	38	41	50
3	51	60	63
4	48	53	49
5	62	55	54

要求根据这些数据推算 100 家企业该指标的总量, 并给出估计的 95% 置信区间。

解: 对这个问题, 我们可以利用二阶抽样的思路解决。首先将企业作为初级单元, 将每一天看做二级单元, 每个企业在调查月内都拥有 30 天 (即拥有 30 个二级单元)。

在这个问题中, 调查人员首先在初级单元中抽取了一个 $n = 5$ 的简单随机样本, 然后对每个样本的二级单元分别独立抽取了一个 $m = 3$ 的简单随机样本, 这就是初级单元大小相等的二阶抽样问题。

由题意, $N = 100, M = 30, n = 5, m = 3$

$$f_1 = \frac{n}{N} = \frac{5}{100} = 0.05, f_2 = \frac{m}{M} = \frac{3}{30} = 0.10$$

首先计算样本初级单元的均值 \bar{y}_i 、方差 s_{2i}^2 :

样本企业	\bar{y}_i	s_{2i}^2
1	60	13
2	43	39
3	58	39
4	50	7
5	57	19

于是得到:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{5} (60 + 43 + 58 + 50 + 57) = 53.6$$

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = 49.3$$

$$s_2^2 = \frac{1}{n} \sum_{i=1}^n s_{2i}^2 = 23.4$$

$$\begin{aligned} v(\bar{y}) &= \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{nm} s_2^2 \\ &= \frac{1-0.05}{5} \times 49.3 + \frac{0.05(1-0.10)}{5 \times 3} \times 23.4 \\ &= 9.3670 + 0.0702 = 9.4372 \end{aligned}$$

计算 \hat{Y} 及 $v(\hat{Y})$:

$$\hat{Y} = NM\bar{y} = 100 \times 30 \times 53.6 = 160800$$

$$v(\hat{Y}) = N^2 M^2 v(\bar{y}) = 100^2 \times 30^2 \times 9.4372 = 84934800$$

\hat{Y} 的标准差为:

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} = \sqrt{84934800} \approx 9216.0078$$

在置信度 95% 的条件下, 对应的 $t = 1.96$, 因此, \hat{Y} 的置信区间为:

$$160800 \pm 1.96 \times 9216$$

或者说在 142736.6 ~ 178863.4 之间。

值得注意的是, 如例 8.1 所示, 方差估计式中, 第一项是主要的, 第二项要小得

多,这是因为第二项的分母是第一项的 m 倍,而且它还要乘以小于1的 f_1 。在最终样本量 $n \times m$ 确定条件下,提高 n 而减小 m 可以大大提高估计的精度。

如果第一阶的抽样比 f_1 可以忽略,则方差估计式(8.11)可以简单为如下的结果:

$$v(\bar{y}) = \frac{s_1^2}{n} = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (8.12)$$

这个结果在实际工作中可以作为参考,因为第二阶抽样采用等距抽样或某些复杂抽样时,方差 S_2^2 的无偏估计很难得到,当 f_1 可以忽略时,只需要初级单元的均值 \bar{y}_i 就可以得到方差近似估计。当然,从另一个方面看, f_1 可以忽略,意味着总体中初级单元 N 很大而抽选出的 n 却很小,结果是样本分布相对集中,势必增大抽样误差。

(二) 对总体比例的估计

欲调查居民户进行家庭装潢时聘请专业装潢公司的比例,这时小区内所有的家庭(二级单元)可以按是否聘请专业装潢公司划分为两类。

如果要估计总体中具有所研究特征的二级单元数占全体二级单元数的比例,则

$$P = \frac{1}{N} \sum_{i=1}^N P_i = \frac{1}{NM} \sum_{i=1}^N A_i$$

式中, A_i 为第 i 个初级单元中具有所研究特征的二级单元数。则对 P 的估计为:

$$p = \frac{1}{n} \sum_{i=1}^n p_i = \frac{1}{nm} \sum_{i=1}^n a_i \quad (8.13)$$

式中, a_i 为第 i 个初级单元中具有所研究特征的二级单元数。

性质3 对于二阶抽样,如果两个阶段都是简单随机抽样,则有

$$E(p) = P$$

估计量 p 的方差为:

$$V(p) = \frac{1}{n} \frac{f_1}{N-1} \sum_{i=1}^N (P_i - P)^2 + \frac{1-f_2}{nm} \frac{M}{N(M-1)} \sum_{i=1}^N P_i Q_i \quad (8.14)$$

$V(p)$ 的无偏估计为:

$$v(p) = \frac{1}{n(n-1)} \sum_{i=1}^n (p_i - p)^2 + \frac{f_1(1-f_2)}{n^2(m-1)} \sum_{i=1}^n p_i q_i \quad (8.15)$$

式中, $Q_i = 1 - P_i$; $q_i = 1 - p_i$ 。

【例8.2】 欲调查某个新小区居民户家庭装潢聘请专业装潢公司的比例。我

们在 15 个单元中随机抽取了 5 个单元,在这 5 个单元中分别随机抽取了 4 户居民并进行了调查,对这 20 户的调查结果如表 8.3 所示。

表 8.3 对 20 个样本户的调查结果

样本单元	第一户	第二户	第三户	第四户
一栋 A 座	是	是	否	否
一栋 C 座	否	是	否	否
二栋 C 座	否	否	否	是
四栋 C 座	否	否	否	否
五栋 B 座	是	否	否	否

要求根据这些数据推算居民家庭装潢聘请专业装潢公司的比例。

解:记请专业装潢公司的居民户为“1”,否则记为“0”。

这里, $N = 15, M = 12, n = 5, m = 4, f_1 = \frac{5}{15}, f_2 = \frac{4}{12}$

因此,聘请专业装潢公司的比例为:

$$p = \frac{1}{nm} \sum_{i=1}^n a_i = \frac{1}{5 \times 4} (2 + 1 + 1 + 0 + 1) = \frac{1}{4} = 0.25$$

其方差的估计是:

$$\begin{aligned}
 v(p) &= \frac{1-f_1}{n(n-1)} \sum_{i=1}^n (p_i - p)^2 + \frac{f_1(1-f_2)}{n^2(m-1)} \sum_{i=1}^n p_i q_i \\
 &= \frac{1 - \frac{5}{15}}{5(5-1)} \left[\left(\frac{2}{4} - \frac{1}{4} \right)^2 + \left(\frac{1}{4} - \frac{1}{4} \right)^2 + \left(\frac{1}{4} - \frac{1}{4} \right)^2 \right. \\
 &\quad \left. + \left(\frac{0}{4} - \frac{1}{4} \right)^2 + \left(\frac{1}{4} - \frac{1}{4} \right)^2 \right] \\
 &\quad + \frac{\frac{5}{15} \left(1 - \frac{4}{12} \right)}{5^2(4-1)} \left(\frac{2}{4} \times \frac{2}{4} + \frac{1}{4} \times \frac{3}{4} + \frac{1}{4} \times \frac{3}{4} + \frac{0}{4} \times \frac{4}{4} + \frac{1}{4} \times \frac{3}{4} \right) \\
 &\approx 0.00657
 \end{aligned}$$

其标准差为: $s(p) = \sqrt{v(p)} \approx 0.081$

因此,可以以 95% 的把握认为,居民装潢请专业公司的比例在

$$0.25 \pm 1.96 \times 0.081$$

的范围内,即 9.1% ~ 40.9% 之间。

§ 8.3 初级单元大小不等的二阶抽样

一般来说,初级单元的大小是不相等的。对于初级单元中的二级单元数不相等的情况,可以通过分层,将大小近似的初级单元分到一层,则层内的二阶抽样就可以按上节介绍的方法来处理

如果按初级单元的大小分层后,层内初级单元的大小差别仍很大,或者合理的分层是按其他指标进行的,则需用到本节介绍的方法来处理二阶抽样的问题。当初级单元大小不等时,对初级单元抽样一般采用不等概抽样。

一、符号说明

首先对初级单元大小不等时二阶抽样使用的符号进行规定。

总体中初级单元个数以及第一阶抽取的样本量: N, n

第 i 个初级单元中二级单元数: M_i

第 i 个初级单元中第二阶抽样的样本量: m_i

第 i 个初级单元中的第 j 个二级单元的观测值: Y_{ij}

样本中第 i 个初级单元中的第 j 个二级单元的观测值: y_{ij}

第一阶和第二阶的抽样比: $f_1 = \frac{n}{N}, f_{2i} = \frac{m_i}{M_i}$

二级单元个数: $M_0 = \sum_{i=1}^N M_i, m_0 = \sum_{i=1}^n m_i$

指标总和: $Y = \sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij}, y = \sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij}$

第 i 个初级单元指标总和: $Y_i = \sum_{j=1}^{M_i} Y_{ij}, y_i = \sum_{j=1}^{m_i} y_{ij}$

第 i 个初级单元按二级单元的平均值: $Y_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij} = \frac{Y_i}{M_i}$

$$y_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} = \frac{y_i}{m_i}$$

$$\text{按二级单元的平均值: } \bar{Y} = \frac{1}{M_0} \sum_{i=1}^n \sum_{j=1}^{M_i} Y_{ij} = \frac{Y}{M_0}, \bar{y} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij}}{\sum_{i=1}^n m_i} = \frac{y}{\sum_{i=1}^n m_i}$$

$$\text{初级单元间的方差: } S_1^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2, s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{y})^2$$

$$\text{第 } i \text{ 个初级单元二级单元间的方差: } S_{2i}^2 = \frac{1}{M_i-1} \sum_{j=1}^{M_i} (Y_{ij} - Y_i)^2$$

$$s_{2i}^2 = \frac{1}{m_i-1} \sum_{j=1}^{m_i} (y_{ij} - y_i)^2$$

二、估计量及其性质

(一) 对初级单元进行简单随机抽样

如果二阶抽样中每个阶段都采用简单随机抽样,并且每个初级单元中二级单元的抽样是相互独立的,则对总体总和的估计可以采用简单估计,也可以考虑采用比率估计。

1. 简单估计量。直观地看,对两个阶段都采用简单随机抽样的二阶抽样,对总体总和的估计可以采用简单估计:

$$\hat{Y}_u = \frac{N}{n} \sum_{i=1}^n M_i y_i = \frac{N}{n} \sum_{i=1}^n \hat{Y}_i \quad (8.16)$$

根据性质 1,不仅可以证明这个估计量是无偏的,并且它的方差为:

$$V(\hat{Y}_u) = \frac{N^2(1-f_1)}{n} \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 + \frac{N}{n} \sum_{i=1}^n \frac{M_i^2(1-f_{2i})}{m_i} S_{2i}^2 \quad (8.17)$$

$V(\hat{Y}_u)$ 的一个无偏估计为:

$$v(\hat{Y}_u) = \frac{N^2(1-f_1)}{n} \frac{1}{n-1} \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}}_u)^2 + \frac{N}{n} \sum_{i=1}^n \frac{M_i^2(1-f_{2i})}{m_i} s_{2i}^2 \quad (8.18)$$

式中,

$$\bar{\hat{Y}}_u = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i \quad (8.19)$$

若一个估计量可以表示成样本观测值总和的常数倍,则称这个样本(或者估计量)是自加权的。对于自加权样本,其估计量的表示形式非常简单,所以在实际工作中,人们通常喜欢将样本构造成为自加权的形式。

由式(8.16)可知,当 $\frac{M_i}{m_i}$ (或者说 $f_{2i} = \frac{m_i}{M_i}$)为常数时,估计量 \hat{Y}_u 就是自加权的。

2. 比率估计量。如果二级单元 Y_{ij} 近似,由于初级单元的大小 M_i 不同,则往往造成初级单元的观测值 Y_i 差异很大,使得估计量方差 $V(\hat{Y}_u)$ 的第一项很大,从而估计量的方差也就变得很大。

这时可以考虑将初级单元的大小 M_i 作为辅助变量,采用比率估计量对总体总和进行估计。

对总体总和的比率估计量为:

$$\hat{Y}_R = M_0 \frac{\sum_{i=1}^n M_i y_i}{\sum_{i=1}^n M_i} = M_0 \frac{\sum_{i=1}^n \hat{Y}_i}{\sum_{i=1}^n M_i} \quad (8.20)$$

这是一个典型的比率估计量,它是有偏的,但随着样本量的增加,其偏倚将趋于零。其近似均方误差为:

$$\begin{aligned} \text{MSE}(\hat{Y}_R) \approx & \frac{N^2(1-f_1)}{n} \frac{1}{N-1} \sum_{i=1}^N M_i^2 (\bar{Y}_i - \bar{Y})^2 \\ & + \frac{N}{n} \sum_{i=1}^N \frac{M_i^2(1-f_{2i})}{m_i} S_{2i}^2 \end{aligned} \quad (8.21)$$

$\text{MSE}(\hat{Y}_R)$ 的样本估计为:

$$\begin{aligned} v(\hat{Y}_R) = & \frac{N^2(1-f_1)}{n} \frac{1}{n-1} \sum_{i=1}^n M_i^2 (\bar{y}_i - \hat{\bar{Y}}_R)^2 \\ & + \frac{N}{n} \sum_{i=1}^n \frac{M_i^2(1-f_{2i})}{m_i} s_{2i}^2 \end{aligned} \quad (8.22)$$

式中,

$$\hat{\bar{Y}}_R = \frac{\hat{Y}_R}{M_0} = \frac{\sum_{i=1}^n M_i y_i}{\sum_{i=1}^n M_i} \quad (8.23)$$

(二) 对初级单元进行放回不等概抽样

对初级单元进行放回不等概抽样时,可以利用第5章介绍的方法,对初级单元进行抽样,即事先规定每个初级单元被抽中的概率 Z_i ($\sum_{i=1}^N Z_i = 1$)。对被抽中的初级单元,再抽取 m_i 个二级单元。如果某个初级单元被抽中多次,则将这 m_i 个二级

单元放回,重新抽取 m_i 个。例如,某个初级单元被重复抽中两次,则对其二级单元抽取一个大小为 m_i 的样本,将这 m_i 个二级单元放回,重新抽取一个大小为 m_i 的样本。当然,这两个样本中的二级单元可能会有重复,应记录下这些样本的情况。实际调查时,对重复的二级单元只调查一次,但计算时,它归哪个样本就参与哪个样本的计算。

对总体总和的估计通常是构造初级单元指标总量 Y_i 的无偏估计 \hat{Y}_i ,然后利用第5章介绍的 Hansen-Hurwitz 估计量对总体总和 Y 进行估计:

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{Y}_i}{z_i} \quad (8.24)$$

由于 \hat{Y}_i 是 Y_i 的无偏估计,由性质1,可以证明 \hat{Y}_{HH} 是 Y 的无偏估计,且 \hat{Y}_{HH} 的方差为:

$$V(\hat{Y}_{HH}) = \frac{1}{n} \left[\sum_{i=1}^n Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2 + \sum_{i=1}^n \frac{V_2(\hat{Y}_i)}{Z_i} \right] \quad (8.25)$$

$V(\hat{Y}_{HH})$ 的一个无偏估计为:

$$v(\hat{Y}_{HH}) = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n \left(\frac{\hat{y}_i}{z_i} - \hat{Y}_{HH} \right)^2 \quad (8.26)$$

注意上述对第二阶抽样并没有做出特别的规定,而且估计量的方差估计式与第二阶抽样的方式无关。

如果希望 \hat{Y}_{HH} 是自加权的,由

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{Y}_i}{z_i} = \frac{1}{n} \sum_{i=1}^n \frac{M_i y_i}{z_i} = \frac{1}{n} \sum_{i=1}^n \frac{M_i}{z_i m_i} \sum_{j=1}^{m_i} y_{ij}$$

则要求

$$\frac{M_i}{nz_i m_i} = K = \frac{1}{f_0} \quad (8.27)$$

这里 f_0 为总体中任意一个二级单元被抽中的概率。如果 f_0 事先确定,则

$$f_{2i} = \frac{m_i}{M_i} = \frac{f_0}{nz_i} \quad (8.28)$$

记总体中所有的二级单元数为 M_0 ,如果抽样时每个初级单元被抽中的概率与其拥有的二级单元数成比例,即初级单元被抽中的概率为 $Z_i = \frac{M_i}{M_0}$,第二阶段对二级单元进行简单随机抽样,则 $m_i = m$ 时,样本是自加权的。这时,对总体总量 Y 的估计为:

$$\hat{Y}_{PPS} = M_0 \bar{y} = \frac{M_0}{n} \sum_{i=1}^n y_i = \frac{M_0}{nm} \sum_{i=1}^n \sum_{j=1}^m y_{ij} \quad (8.29)$$

估计量的方差估计为:

$$v(\hat{Y}_{PPS}) = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (8.30)$$

实际工作中,如果初级单元大小不相等,通常人们喜欢在第一阶抽样时按放回的与二级单元数成比例的不等概抽样,第二阶抽样则进行简单随机抽样,且每个样本初级单元的样本量都相同,这样得到的样本是自加权的,估计量的形式非常简单。

【例 8.3】 某小区拥有 10 座高层建筑,每座高层建筑拥有的楼层数如表 8.4 所示。

表 8.4 10 座高层建筑的各自层数

高层建筑	A	B	C	D	E	F	G	H	I	J
楼层	12	12	16	15	10	16	10	18	16	20

用二阶抽样方法抽出 10 个楼层进行调查,第一阶抽样为放回的、按与每座建筑拥有的楼层数成比例的不等概抽样抽取 5 座建筑,第二阶按简单随机抽样对每座建筑抽取两个楼层。对 10 个楼层居民人数的调查结果如表 8.5 所示,请对小区总居民数进行估计,并给出估计的精度。

表 8.5 中选的一阶样本序号和 10 个楼层的居民数

阶样本序号	1	2	3	4	5
居民数	18, 12	15, 18	19, 13	16, 10	16, 11

解:已知 $n = 5, m = 2, M_0 = 145, \sum_{i=1}^n \sum_{j=1}^m y_{ij} = 148$

注意到这个样本是自加权的,根据公式(8.29),得

$$\hat{Y} = \frac{M_0}{nm} \sum_{i=1}^n \sum_{j=1}^m y_{ij} = \frac{145}{5 \times 2} \times 148 = 2146(\text{人})$$

$$\bar{y} = \frac{\hat{Y}}{M_0} = \frac{2146}{145} = 14.8$$

估计量的方差:

$$v(\hat{Y}) = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$= \frac{145^2}{5(5-1)} [(15-14.8)^2 + (16.5-14.8)^2 + (16-14.8)^2 + (13-14.8)^2 + (13.5-14.8)^2] \\ = 9776.625$$

估计量的标准差为:

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} \approx 98.88$$

因此,小区居民数为 2146 人,在置信度为 95% 时,估计的相对误差为:

$$r = \frac{s(\hat{Y})}{\hat{Y}} = \frac{1.96 \times 98.88}{2146} \approx 9\%$$

(三) 对初级单元进行不放回不等概抽样

不放回不等概抽样的效率比放回的效率要高,因此,有时人们也会倾向于用不放回不等概抽样来抽取初级单元,通过简单随机抽样获得二级单元。这时可以利用第 5 章介绍的不放回不等概抽样的结果对总体总量进行推算。同第 5 章介绍的情形一样,这时估计量的推算比较复杂。

如果初级单元的包含概率为 π_i 及 π_{ij} , 对总体总量 Y 的估计可以采用 Horvitz-Thompson(霍维茨-汤普森)估计:

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{\hat{Y}_i}{\pi_i} = \sum_{i=1}^n \frac{M_{iy_i}}{\pi_i} \quad (8.31)$$

\hat{Y}_{HT} 方差的估计为:

$$v(\hat{Y}_{HT}) = \sum_{i=1}^n \frac{1-\pi_i}{\pi_i^2} \hat{Y}_i^2 + 2 \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \hat{Y}_i \hat{Y}_j + \sum_{i=1}^n \frac{M_i^2 (1-f_{2i})}{m_i \pi_i} s_{2i}^2 \quad (8.32)$$

如果 n 固定,则 $V(\hat{Y}_{HT})$ 也可用

$$v(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j} \right)^2 + \sum_{i=1}^n \frac{M_i^2 (1-f_{2i})}{m_i \pi_i} s_{2i}^2 \quad (8.33)$$

来估计。

§ 8.4 其他问题

一、总样本量及最优样本量的配置

对于二阶抽样,应该抽多少二级单元,即确定 nm 为多少,可以采用两种办法。一种是根据调查费用,确定可以调查的样本量;另一种是根据简单随机抽样时应抽样本量,再乘以设计效应 $deff$ 获得。二阶抽样与简单随机抽样相比,其效率比较低, $deff$ 应该大于 1。实际工作中,对于二阶抽样,有时取 $deff$ 的经验数据(不同项目的 $deff$ 不同,案例分析的资料显示, $deff$ 的范围约在 1.3 ~ 3 之间)。

对于初级单元大小相等的二阶抽样,如何设计两个阶段的样本量,即如何确定 n 和 m 是需要考虑的问题。由于影响精度的主要原因是初级单元之间的差异,所以多抽一些初级单元,少抽一些二级单元比较好,但往往初级单元的调查费用比二级单元要高。好的设计可以在调查总费用一定的情况下,使估计的精度最高;或在一定的精度条件下,使调查总费用最省,这就是最优样本量的配置或最优抽样比 f_1 和 f_2 的确定问题。

考虑费用函数为最简单的一种形式为:

$$C = c_0 + c_1 n + c_2 nm \quad (8.34)$$

式中, c_0 为与样本量无关的固定费用,如公司的办公费、场租费等; c_1 为每调查一个初级单元的费用; c_2 为每调查一个二级单元的费用。

则 m 的最优值为:

$$m_{opt} = \frac{S_2}{S_u} \sqrt{\frac{c_1}{c_2}} \quad (8.35)$$

式中, $S_u^2 = S_1^2 - \frac{S_2^2}{M}$

实际使用时, m 应为整数,但计算出的 m_{opt} 往往不是整数,令 m' 为 m_{opt} 的整数部分,则 m 的取值规则为:

- (1) 当 $m_{opt}^2 > m'(m' + 1)$, 则取 $m = m' + 1$;
- (2) 当 $m_{opt}^2 \leq m'(m' + 1)$, 则取 $m = m'$;
- (3) 当 $m_{opt}^2 > M$ 或 $S_1^2 - \frac{S_2^2}{M} < 0$, 则取 $m = M$ 。

求出 m 之后,根据总费用函数,就可以确定 n ,从而确定最优抽样比 f_1 和 f_2 。

【例 8.4】(续例 8.1) 若 $\frac{c_1}{c_2} = 10, V(\bar{y}) = 15$, 试确定最优 m, n

解: 首先计算 m_{opt}

由例 8.1 的计算, 知

$$s_1^2 = 49.3, s_2^2 = 23.4$$

于是, 由本章附录 2, 有

$$\hat{S}_1^2 = s_1^2 - \frac{1-f_2}{m}s_2^2 = 49.3 - \frac{1-0.1}{3} \times 23.4 = 42.28$$

$$\hat{S}_2^2 = s_2^2 = 23.4$$

$$\hat{S}_u^2 = \hat{S}_1^2 \frac{\hat{S}_2^2}{M} = 42.28 \frac{23.4}{30} = 41.5$$

因此

$$m_{opt} = \frac{\hat{S}_2}{\hat{S}_u} \sqrt{\frac{c_1}{c_2}} = \sqrt{\frac{23.4}{41.5}} \times 10 \approx 2.37$$

$$m' = 2, m' + 1 = 3$$

因为 $m_{opt}^2 \approx 5.64 < m'(m' + 1) = 6$

因而取最优的 $m = 2$ 。

进一步计算 n_{opt} 。

$$\text{由 } V(\bar{y}) = \frac{1}{n} \left(S_1^2 - \frac{S_2^2}{M} \right) + \frac{S_2^2}{nm} = \frac{S_1^2}{N}$$

因此

$$15 = \frac{1}{n_{opt}} \left(42.28 - \frac{23.4}{30} \right) + \frac{23.4}{n_{opt} \times 2} = \frac{42.28}{100}$$

整理得到

$$n_{opt} \approx 3.449$$

因而可以取 $n = 4$ 。

二、三阶及多阶段抽样

(一) 各级单元大小相等时的多阶段抽样

二阶抽样的推广是三阶段抽样, 乃至更高阶抽样。对于三阶段抽样, 前两阶与二阶抽样相同, 只是第三阶段的抽样是对被抽中的二级单元中的三级单元再抽样, 从中抽出样本三级单元(接受调查的最终单元)。

如果总体拥有 N 个初级单元, 每个初级单元拥有 M 个二级单元, 每个二级单元又拥有 K 个三级单元, 各阶的样本量分别为 n, m, k , 每个阶段都按简单随机抽样, 则三级单元总体均值的估计为:

$$\bar{y} = \frac{1}{nmk} \sum_{i=1}^n \sum_{j=1}^m \sum_{u=1}^k y_{iju} \quad (8.36)$$

其方差为:

$$V(\bar{y}) = \frac{1}{n} \frac{f_1}{S_1^2} + \frac{1-f_2}{nm} S_2^2 + \frac{1-f_3}{nmk} S_3^2 \quad (8.37)$$

其无偏估计为:

$$v(\bar{y}) = \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{nm} s_2^2 + \frac{f_1 f_2 (1-f_3)}{nmk} s_3^2 \quad (8.38)$$

对照二阶抽样的估计公式, 可以看出, 对于更高阶的抽样, 对(最终单元的)均值的估计就是样本均值, 也就是将所有最终样本单元的指标求和, 然后除以最终单元的样本量。

由于方差的主要项为第一项, 其次为第二项, 第三项几乎很小了, 所以对于更高阶的抽样, 根据不同的情况(如各阶的样本量, 各阶内单元之间的方差等), 估计量的方差计算一般只计算到第二阶至第三阶就可以了。

(二) 各级单元大小不相等时的多阶段抽样

1. 各阶抽样采用不等概抽样。一般情况下, 各级单元的大小不相等。类似对初级单元大小不等的二阶抽样时的讨论, 通常这时每一阶的抽样采用与单元大小成比例的不等概抽样, 而且通常抽样是放回的, 即 PPS 抽样。

以三阶抽样为例。记:

总体拥有 N 个初级单元, 每个初级单元拥有 M_i 个二级单元, 每个二级单元又拥有 K_{ij} 个三级单元。

各阶样本量分别为 n, m, k (注意 m, k 不随单元变化), 即抽取 n 个初级单元, 在每个样本初级单元中, 抽取 m 个二级单元, 在每个样本二级单元中, 抽取 k 个三级单元

每一阶单元被抽中的概率为 Z_i, Z_{ij}, Z_{iju} , 它们满足:

$$\sum_{i=1}^N Z_i = 1, \sum_{j=1}^{M_i} Z_{ij} = 1, \sum_{u=1}^{K_{ij}} Z_{iju} = 1$$

这时对总体总和

$$Y = \sum_{i=1}^N \sum_{j=1}^{M_i} \sum_{u=1}^{K_{ij}} Y_{iju}$$

的无偏估计为:

$$\hat{Y} = \frac{1}{nmk} \sum_{i=1}^n \frac{1}{z_i} \sum_{j=1}^m \frac{1}{z_{ij}} \sum_{u=1}^k \frac{y_{iju}}{z_{iju}} \quad (8.39)$$

其方差为:

$$V(\hat{Y}) = \frac{1}{n} \left(\sum_{i=1}^n \frac{Y_i^2}{Z_i} - Y^2 \right) + \frac{1}{nm} \sum_{i=1}^n \frac{1}{Z_i} \left(\sum_{j=1}^{M_i} \frac{Y_{ij}^2}{Z_{ij}} - Y_i^2 \right) + \frac{1}{nmk} \sum_{i=1}^n \frac{1}{Z_i} \left[\sum_{j=1}^{M_i} \frac{1}{Z_{ij}} \left(\sum_{u=1}^{K_{ij}} \frac{Y_{iju}^2}{Z_{iju}} - Y_{ij}^2 \right) \right] \quad (8.40)$$

式中,

$$Y_{ij} = \sum_{u=1}^{K_{ij}} Y_{iju}, Y_i = \sum_{j=1}^{M_i} \sum_{u=1}^{K_{ij}} Y_{iju} = \sum_{j=1}^{M_i} Y_{ij} \quad (8.41)$$

$V(\hat{Y})$ 的一个无偏估计为:

$$v(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^n (\hat{Y}_i - \hat{Y})^2 \quad (8.42)$$

式中,

$$\hat{Y}_i = \frac{1}{z_i m} \sum_{j=1}^m \frac{1}{z_{ij}} \left(\frac{1}{k} \sum_{u=1}^k \frac{y_{iju}}{z_{iju}} \right)$$

2. 样本为自加权的条件。实际工作中,通常的做法是前两阶抽样采用 PPS 抽样,即对初级单元和二级单元的抽样按放回的、与其单元大小成比例的概率抽样;最后一阶抽样按等概率抽选。如果从第二阶开始,每一阶的样本量都相同(即 $m_i = m, k_j = k$),则样本是自加权的。

这时,

$$Z_i = \frac{\sum_{j=1}^{M_i} K_{ij}}{\sum_{i=1}^N \sum_{j=1}^{M_i} K_{ij}} = \frac{\sum_{j=1}^M K_{ij}}{M_0}, Z_{ij} = \frac{K_{ij}}{\sum_{j=1}^{M_i} K_{ij}}, Z_{iju} = \frac{1}{K_{ij}}$$

注意这时第三阶抽样也是放回的,各阶单元的大小是以最小(最终)单元数计算的。

将其代入式(8.39),则总体总和的估计为:

$$\hat{Y} = \frac{M_0}{nmk} \sum_{i=1}^n \sum_{j=1}^m \sum_{u=1}^k y_{iju} = M_0 \bar{y} \quad (8.43)$$

\bar{y} 是以三级单元计算的样本简单平均数。 \hat{Y} 的表达式正好说明这时估计量是自加权的。

\hat{Y} 方差的估计为:

$$v(\hat{Y}) = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2 \quad (8.44)$$

式中, $\bar{\bar{y}} = \frac{1}{mk} \sum_{j=1}^m \sum_{u=1}^k y_{ju}$

如果对一级单元的抽样采用不放回简单随机抽样,上述公式仍然成立,只是估计量的理论方差比放回的情形小一些。

类似地,对于更高阶的情形,除了最后一阶采用等概率抽样(放回的或不放回的均可),前几阶均采用 PPS 抽样,并且自第二阶开始,每一阶的样本量都相同(即 $m_i = m, k_j = k, \dots$),则样本是自加权的,其估计量的形式非常简单。

【例 8.5】 某调查公司接受了一项关于全国城市成年居民人均奶制品消费支出及每天至少喝一杯鲜奶的人数的比例情况的调查。确定抽样范围为全国地级及以上城市中的成年居民。成年居民指年满 18 周岁以上的居民。

第一步:确定抽样方法。

调查公司决定采用多阶段抽样方法进行方案设计,调查的最小单元为成年居民。确定调查的各个阶段为城市、街道、居委会、居民户,在居民户中利用二维随机表(Kish 随机表的简化)抽取成年居民。

第二步:确定样本量及各阶段样本量的配置。

按简单随机抽样时,在 95% 置信度下,绝对误差为 5%,取使方差达到最大的(消费奶制品的居民)比例 50%,则全国样本量应为:

$$n_0 = \frac{t^2 PQ}{d^2} \approx \frac{2^2 \times 0.5 \times 0.5}{0.05^2} = 400(\text{人})$$

根据以往调查的经验,估计回答率 $b = 80\%$,因此调整样本量为:

$$n_1 = \frac{n_0}{b} = \frac{400}{0.8} = 500(\text{人})$$

多阶段抽样的效率比简单随机抽样的效率低,这里取设计效应 $deff = 3.2$,则在全国范围内应调查的样本居民为:

$$n_2 = n_0 \times deff = 400 \times 3.2 = 1280(\text{人})$$

各阶段的样本量配置为:

初级单元:20 个城市;

二级单元:80 个街道,每个样本市内抽 4 个街道;

三级单元:160 个居委会,每个样本街道内抽 2 个居委会;

四级单元:1280 个居民户,每个样本居委会内抽 10 个居民户。

在样本居民户内,利用二维随机表抽1名成年居民。

第二步:确定抽样方法。

第一阶段,在全国城市中按与人口数成比例的放回的不等概抽样,即PPS抽样。

第二阶段和第三阶段分别按与人口数成比例的不等概等距抽样。

以第二阶段为例,在某个被抽中的样本城市中,将其所属的街道编号,搜集各街道的人口数,赋予每个街道与其人口相同的代码数;根据该市总人口数除以样本量4,确定抽样间距;然后对代码进行随机起点的等距抽样,则被抽中代码所在的街道为样本街道。

第四阶段,分别在每个样本居委会中,按等距抽样抽出10个居民户。即根据居委会拥有的居民户数除以样本量10得到抽样间距,然后随机起点等距抽样。

在每个样本居民户中,调查员按二维随机表抽取1名成年居民。二维随机表的使用方法如下。

(1) 随机号的确定。应事先在随机表的第一行数字上,选好一个数字,并划上一个圈,被圈好的这个数字就是这份问卷的随机号。随机号的选择一般由小到大或循环给出。可以根据便于操作又保证实现随机的原则,选择确定随机号的适当方法。

(2) 选出被访者。将所有符合基本要求的家庭成员按年龄从大到小的顺序列入随机表中,以事先做好的随机号为纵坐标、以最小家庭成员为横坐标,交叉处对应的数字即为被访者的序号。例如,某受访户的随机号确定为4,该户中家庭成员符合本次调查要求的共有4人。将这4人的基本情况按年龄从大到小的顺序填入下面的随机表中。如表8.6所示。

表 8.6

序号	姓名	性别	年龄	1	2	3	④	5	6	7	8	9	10	11	12
1	肖明	男	53	1	1	1	1	1	1	1	1	1	1	1	1
2	汪红	女	52	2	1	1	2	1	2	1	2	1	2	2	1
3	肖晓波	男	23	3	2	1	1	3	2	2	1	3	1	2	3
4	肖晓玲	女	21	4	1	3	②	2	3	1	4	3	2	4	1
5				5	4	1	2	3	4	1	2	3	5	4	2
6				6	3	1	5	2	4	3	5	1	4	6	2
7				7	1	4	3	6	2	5	2	5	7	4	3
8				8	4	5	7	1	2	6	3	7	5	3	1
9				9	5	1	4	3	8	2	7	6	5	2	8
10				10	3	5	9	4	1	7	2	8	6	9	4
11				11	6	1	5	10	4	9	8	3	2	7	6
12				12	7	2	9	4	11	6	1	8	3	10	5

表中,序号为4的列与年龄最小的家庭成员肖晓玲所在的第4行交叉的数字是2。因此,第2号家庭成员汪红为被访者。

第四步:推算方法。

这样获得的样本,虽然不是严格按照前四阶采用PPS抽样、最后一阶采用等概率抽样,但由于每一阶的抽样比相对来说可以忽略,因此它仍可以近似地作为一个自加权样本,这时,可以将样本均值作为总体均值的无偏估计。

记各样本城市的80位样本居民中,奶制品消费总支出为 y_i ,则各样本城市人均奶制品消费支出为:

$$y_i = \frac{y_i}{80}, i = 1, \dots, n$$

全国1600名居民组成的样本中,奶制品消费总支出为 $y = \sum_{i=1}^n y_i$,则成年居民人均奶制品消费支出为:

$$\bar{y} = \frac{y}{1600} = \frac{1}{1600} \sum_{i=1}^n y_i$$

y 的方差的估计为:

$$v(y) = \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2$$

对总体比例的推算可以借用对均值的推算公式。记各样本城市的80位样本居民中,每天至少喝一杯鲜奶的人数为 a_i ,则各样本城市每天至少喝一杯鲜奶的人数的比例为:

$$p_i = \frac{a_i}{80}, i = 1, \dots, n$$

全国1600名居民组成的样本中,每天至少喝一杯鲜奶的总人数为 $a = \sum_{i=1}^n a_i$,则成年居民中每天至少喝一杯鲜奶的人数的比例为:

$$\bar{p} = \frac{a}{1600} = \frac{1}{1600} \sum_{i=1}^n a_i$$

p 的方差的估计为:

$$v(p) = \frac{1}{n(n-1)} \sum_{i=1}^n (p_i - \bar{p})^2$$

以上公式中 $n = 20$ 。

小 结

本章介绍了多阶段抽样方法。对于大规模的抽样调查项目,通常采用多阶段抽样方法。这种方法可以看做对样本群内的单元进行再抽样的一种方法,和整群抽样情形一样,当各级单元大小相同时,各阶的抽样采用等概率抽样的方法。但实际中,大多数是各级单元大小不等的情形,这时,最简单的方法是构造自加权的样本,也就是前几阶采用 PPS 抽样,最后一阶采用等概率抽样,并且从第二阶开始,每一阶的样本量都相同,这时估计量的形式非常简单。

本章附录 多阶段抽样估计量性质的证明

1. 证明性质 1。

证明:这里只给出两阶段抽样时估计量均值、方差的计算公式,三阶段抽样时的公式推导类似。

对于均值公式

$$E(\hat{\theta}) = E_1 E_2(\hat{\theta})$$

可以理解为对所有可能样本的平均,可以分两步进行。在给定的一个样本量为 n 的初级单元样本中,对所有二级抽样可能的样本估计量进行平均,然后再对所有一级抽样可能的样本估计量进行平均。

$$\text{记 } E(\hat{\theta}) = \bar{\theta}$$

$$V(\hat{\theta}) = E(\hat{\theta} - \bar{\theta})^2 = E_1 E_2(\hat{\theta} - \bar{\theta})^2$$

由

$$\begin{aligned} E_2(\hat{\theta} - \bar{\theta})^2 &= E_2(\hat{\theta})^2 - 2\bar{\theta}E_2(\hat{\theta}) + \bar{\theta}^2 \\ &= [E_2(\hat{\theta})]^2 + V_2(\hat{\theta}) - 2\bar{\theta}E_2(\hat{\theta}) + \bar{\theta}^2 \end{aligned}$$

对两边求 E_1 , 得

$$\begin{aligned} V(\hat{\theta}) &= E_1[E_2(\hat{\theta})]^2 + E_1[V_2(\hat{\theta})] - \bar{\theta}^2 \\ &= E_1[E_2(\hat{\theta})]^2 + E_1[V_2(\hat{\theta})] - [E_1 E_2(\hat{\theta})]^2 \end{aligned}$$

$$V_1[E_2(\hat{\theta})] + E_1[V_2(\hat{\theta})]$$

2. 证明性质 2: \bar{Y} 的无偏估计及其方差。

证明: 要证明 \bar{y} 是 \bar{Y} 的无偏估计, 需要用到性质 1。

$$E(\bar{y}) = E_1 E_2(\bar{y})$$

由于两个阶段的抽样都是简单随机的, 因此由简单随机抽样的性质, 有

$$\begin{aligned} E(\bar{y}) &= E_1[E_2(\bar{y})] = E_1\left[E_2\left(\frac{1}{n} \sum_{i=1}^n y_i\right)\right] = E_1\left[\frac{1}{n} \sum_{i=1}^n E_2(y_i)\right] \\ &= E_1\left(\frac{1}{n} \sum_{i=1}^n \bar{Y}_i\right) = \bar{Y} \end{aligned}$$

由于每个初级单元中对二级单元的抽样是相互独立的, 因此 \bar{y} 的方差

$$V(\bar{y}) = V_1[E_2(\bar{y})] + E_1[V_2(\bar{y})]$$

$V(\bar{y})$ 的第一项

$$\begin{aligned} V_1[E_2(\bar{y})] &= V_1\left[E_2\left(\frac{1}{n} \sum_{i=1}^n y_i\right)\right] = V_1\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \\ &= V_1\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1-f_1}{n} \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2 \\ &= \frac{1-f_1}{n} S_1^2 \end{aligned}$$

$V(\bar{y})$ 的第二项

$$\begin{aligned} E_1[V_2(\bar{y})] &= E_1\left[V_2\left(\frac{1}{n} \sum_{i=1}^n \bar{y}_i\right)\right] = E_1\left[\frac{1}{n^2} \sum_{i=1}^n V_2(y_i)\right] \\ &= E_1\left[\frac{1}{n^2} \sum_{i=1}^n \left[\frac{1-f_2}{m} \frac{1}{M-1} \sum_{j=1}^M (Y_{ij} - Y_i)^2\right]\right] \\ &= E_1\left[\frac{1}{n^2} \sum_{i=1}^n \left(\frac{1-f_2}{m} S_{2i}^2\right)\right] \\ &= \frac{1-f_2}{nm} E_1\left(\frac{1}{n} \sum_{i=1}^n S_{2i}^2\right) = \frac{1-f_2}{nm} \left(\frac{1}{N} \sum_{i=1}^N S_{2i}^2\right) = \frac{1-f_2}{nm} S_2^2 \end{aligned}$$

从而得到

$$V(\bar{y}) = \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{nm} S_2^2$$

下面证明 $E[v(\bar{y})] = V(\bar{y})$, 这需要先求得 $E(s_1^2)$ 和 $E(s_2^2)$ 。注意到每个初级单元中二级单元的抽样是相互独立的, 因此有

$$E_2[(n-1)s_1^2] = E_2\left[\sum_{i=1}^n (y_i - \bar{y})^2\right] = \sum_{i=1}^n E_2(y_i^2) - nE_2(\bar{y}^2)$$

$$\begin{aligned}
&= \sum_{i=1}^n [E_2(y_i)]^2 + V_2(y_i) = n [E_2(\bar{y})]^2 + V_2(\bar{y}) \\
&= \sum_{i=1}^n \left(Y_i^2 + \frac{1-f_2}{m} S_{2i}^2 \right) - n \left(\frac{1}{n} \sum_{i=1}^n Y_i \right)^2 - \\
&\quad \frac{1-f_2}{nm} \sum_{i=1}^n S_{2i}^2
\end{aligned}$$

记 $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$ (注意 $\bar{Y}_n \neq \bar{Y}$), 则

$$\begin{aligned}
E_2[(n-1)s_1^2] &= \sum_{i=1}^n (Y_i^2) - n(\bar{Y}_n)^2 + \frac{1-f_2}{m} \sum_{i=1}^n S_{2i}^2 - \frac{1-f_2}{nm} \sum_{i=1}^n S_{2i}^2 \\
&= \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 + \frac{(n-1)(1-f_2)}{nm} \sum_{i=1}^n S_{2i}^2
\end{aligned}$$

于是有

$$\begin{aligned}
E(s_1^2) &= E_1[E_2(s_1^2)] \\
&= E_1\left[\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2\right] + \frac{1-f_2}{m} E_1\left(\frac{1}{n} \sum_{i=1}^n S_{2i}^2\right) \\
&= S_1^2 + \frac{1-f_2}{m} S_2^2
\end{aligned}$$

对于 s_2^2 , 有

$$\begin{aligned}
E(s_2^2) &= E_1[E_1(s_2^2)] = E_1\left\{E_2\left[\frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - y_i)^2\right]\right\} \\
&= E_1\left\{\frac{1}{n} \sum_{i=1}^n E_2\left[\frac{1}{m-1} \sum_{j=1}^m (y_{ij} - y_i)^2\right]\right\} \\
&= E_1\left[\frac{1}{n} \sum_{i=1}^n \frac{1}{m-1} \sum_{j=1}^m (Y_{ij} - Y_i)^2\right] = E_1\left(\frac{1}{n} \sum_{i=1}^n S_{2i}^2\right) \\
&= \frac{1}{N} \sum_{i=1}^N S_{2i}^2 = S_2^2
\end{aligned}$$

将 $E(s_1^2), E(s_2^2)$ 代入 $E[v(\bar{y})]$, 得

$$\begin{aligned}
E[v(\bar{y})] &= \frac{1}{n} f_1 \left(S_1^2 + \frac{1}{m} \frac{f_2}{m} S_2^2 \right) + \frac{f_1(1-f_2)}{nm} S_2^2 \\
&= \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{nm} S_2^2 = V(\bar{y})
\end{aligned}$$

习 题

1. 一项关于居民情况的调查,调查人员根据手头的一份居民户名单抽中了一批样本居民户,如果调查时不能耽误样本户很多时间,对于以下的调查项目,判断是否有必要对样本居民户中的居民进行再抽样:

- (1) 居民性别比;
- (2) 对甲 A 足球队下次比赛各队名次的预测;
- (3) 人均月用水量;
- (4) 对汽车品牌认知度。

2. 某高校欲利用二阶抽样方法调查下述指标,请你分别选择两个阶段合适的抽样单元,并叙述理由。

- (1) 全校学生拥有的电脑数;
- (2) 为了学习英语,平均每位同学拥有的各种英语教学书籍;
- (3) 测试男生平均每分钟俯卧撑次数。

3. 某高校学生会欲对全校女生拍摄过个人艺术照的比例进行调查。全校共有女生宿舍 200 间,每间住 6 位同学,学生会的同学运用二阶抽样设计了抽样方案,从 200 间宿舍中抽取了 10 间样本宿舍,在每间样本宿舍中抽取了 3 位同学分别进行单独访问,两个阶段的抽样都是简单随机抽样,调查的结果如下:

样本宿舍	拍照人数	样本宿舍	拍照人数
1	2	6	1
2	0	7	0
3	1	8	1
4	2	9	1
5	1	10	0

试估计拍摄过个人艺术照的女生的比例,并给出估计的标准差。

4. 上题中,学生会女生勤工助学月收入的一项调查中,根据以往同类问题的调查,宿舍间的标准差为 $S_1 = 326$ 元,宿舍内同学之间的标准差为 $S_2 = 188$ 元。以一位同学进行调查来计算,调查每个宿舍的时间 c_1 为 10 分钟,调查每一位学生的时间 c_2 为 1 分钟,为了调查需要做各方面的准备及数据计算等工作,所花费的时间是 c_0 为 4 小时,如果总的时间控制在 8 小时内,则最优的样本宿舍和样本学生数

为多少?

5. 某居委会欲了解居民健身活动情况,如果已知该居委会会有500名居民,在所属10个单元中抽出了4个单元,然后在样本单元中分别抽出若干居民,两个阶段的抽样都是简单随机抽样,调查了样本居民每天用于健身活动的时间结果如下(以10分钟为1个单位):

单元(i)	居民人数(M_i)	样本量(m_i)	时间(y_{ij})
1	32	4	4,2,3,6
2	45	5	2,2,4,3,6
3	36	4	3,2,5,8
4	54	6	4,3,6,2,4,6

试估计居民平均每天用于锻炼的时间,并给出估计的标准差。

(1) 简单估计量;

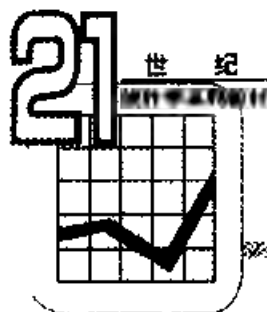
(2) 比率估计量;

(3) 对两种估计方法及估计结果进行评价。

6. 某公司欲了解职工上班交通所需时间,该公司共有5个部门,根据每个部门的人数采用PPS抽样抽出2个部门,并在2个部门中采用简单随机抽样分别抽出5名职工,调查的结果如下:

部门(i)	职工人数(M_i)	时间(y_{ij} , 分钟)
1	20	40,10,20,30,40
2	35	60,30,20,60,30

试估计该公司职工上班交通平均所需时间,并给出估计的标准差。



第 9 章

二重抽样

前面各章介绍的几种抽样技术中,大都需要事先了解一些关于总体的信息,例如分层抽样需要事先知道各层权重,比率估计和回归估计中需要知道总体的某些辅助信息,但在一些情况下,这些资料在调查前无法预知。这时,我们可以先从总体中抽取一个大的初始样本,从而获得总体的辅助信息,然后再从初始样本或从总体中再抽一个子样本,这种方法就是二重抽样。本章第一节介绍二重抽样的定义、作用及其与两阶段抽样的区别,第二节介绍为分层抽样进行的二重抽样,第三节介绍为比率估计进行的二重抽样,第四节介绍为回归估计进行的二重抽样。

§ 9.1 引言

一、定义

二重抽样(double sampling),也称二相抽样或两相抽样(two phase sampling),是指在抽样时分两步抽取样本。一般情况下,先从总体 N 中抽取一个较大的样本 n' ,称为第一重(相)样本(the first phase sample),对之进行调查以获取总体的某些辅助信息,为下一步的抽样估计提供条件;然后进行第二重(相)抽样

(the second phase sampling)。第二重抽样所抽的样本 n 相对较小,但是第二重抽样调查才是主调查。一般地,第二重样本(the second phase sample)是从第一重样本中抽取的,也即第一重样本的子样本,但有时也可以从总体中独立地抽取。由于样本是分两次抽取的,因此称做二重抽样。

例如,欲对某城市体育场馆的营业状况进行抽样调查,鉴于不同场馆功能和面积差异较大,拟采用分层抽样,但由于缺乏分层资料,故先随机抽选一个较大的样本,对该样本仅进行分层及进行层权估计,费用相对较低;然后利用第一次调查获得的分层资料,进行一次较小样本的分层抽样,对该样本进行一次正式调查。这就是二重抽样。

显然,二重抽样方法也可以推广到多次抽取样本,然后结合起来对总体的有关标志值进行估计,这就是多重抽样或多相抽样。本章主要讨论二重抽样。

二、二重抽样与两阶段抽样

二重抽样和两阶段抽样,在名称上很容易引起混淆。虽然二者都可被视为一种分段抽样方法,但是二重抽样和二阶段抽样的差异还是很显著的。首先,两阶段抽样(two-stage sampling)是先从总体 N 个单元(初级单元)中抽出 n 个样本单元,却并不对这 n 个样本单元中的所有小单元(二级单元)都进行调查,而是在其中再抽出若干个二级单元进行调查;二重抽样则不同,要对第一重(相)样本(the first phase sample)进行调查以获取总体的某些辅助信息,并且要利用这些辅助信息进行排序、分层、抽样或估计等。其次,两阶段抽样的第二阶段抽样单元和第一阶段抽样单元往往是不同的,比如第一阶段抽样单元是居委会,第二阶段抽样单元是住户;而二重抽样的第二重样本则往往是第一重样本的子样本,两次抽样的单元是相同的。也就是说,二重抽样要有一份最终单元的完整名册(总体所有单元的抽样框),而两阶段抽样只是需要第一阶段单元(初级单元)名册,然后在中选的初级单元中构造第二阶段抽样的抽样框。

例如,如果某城市想做一次消费调查,只有一份总户册,没有任何分类信息,调查时先取一个住户的大样本调查分层信息,再利用分层信息从中抽取小样本进行详细调查,这是二重抽样。如果某市没有总户册,但有居委会名册,抽样时先抽取居委会,再从居委会中抽取住户,对其进行调查,这是两阶段抽样。

三、二重抽样的作用

(一) 有助于筛选主调查对象

在一些调查中,调查对象只是总体中的一个部分,且与其他单元不易区分。例

如对某品牌化妆品的用户进行入户调查,调查前并不知道该样本是否为调查对象;再如在 一项办公自动化设备调查中,要求调查单元的微机、复印机与传真机等办公自动化设备的使用情况,但事先也不好确定哪些单元一定有这些设备。这时,就可以采用二重抽样,先从总体中抽取一个大样本,通过相对比较简单的调查测试,筛选出满足条件的对象,从中再抽样进行进一步的主调查。

(二) 节约调查费用

对于一项大规模的多指标调查,由于单元之间的差异或对目标量估计的精度要求不同,往往并不需要相同的样本量。例如在城市居民的住户家计调查中,对家用耐用品、旅游开支等指标的调查,要达到一定的精度需要较大的样本量;而对家庭日用品、粮食、油盐酱醋开支等指标的调查,由于其差别较小,因此要达到同样的精度,其样本量就不必很大。这时可以采用二重抽样,先抽取一个大样本,对差异较大的项目或精度要求比较高的项目进行调查,然后再抽一个较小的样本,对差异较小的项目进行调查,则可在保证一定精度的前提下节约调查费用。

(三) 提高抽样效率

许多抽样技术都需要利用已有的辅助信息来提高抽样效率,例如分层随机抽样需要事先将总体单元进行分层,知道层权;比率估计或回归估计则需要知道有关辅助变量的总体总和或均值。然而并非任何时候都能够获得所需要的总体辅助信息,这时采用二重抽样方法,先抽取一个较大的样本以获取有用的信息,然后再抽取一个较小的样本做出改进的估计,就是一个提高抽样效率的可行选择。需要指出的是,在抽取第一重样本时需要增加一定的费用,只有当利用这些信息进行分层抽样,在比率估计和回归估计时提高精度的得益大于所增加的费用时,采用二重抽样才是合算的。

(四) 可用于研究样本轮换中的某些问题

许多调查需要经常性地定时进行,如农产量调查、家计调查等,需要对同一总体进行连续抽样。在连续抽样中,利用连续时间序列样本不同时间的指标值之间的相关性可以提高估计的精度,但是长期使用固定样本单元,则会由于样本疲劳或样本老化的现象而影响调查的质量,这时则可以采用样本轮换(sample rotation)的方法以提高估计精度。在样本轮换问题的研究中二重抽样方法有很好的应用。

(五) 降低无回答偏倚

高无回答率及其递增的趋势一直困扰着调查行业。在对无回答的补救方法中,二重抽样方法受到广泛的注意。这种方法的思想是,对最初的无回答进行再一次的随机抽样,对无回答子样本采用更细致、更艰巨的努力去获得其数据,用第一次样本的回答数据和第二次样本数据进行估计,以消除无回答的偏倚影响,改善对总体

的估计效果。无回答的二级抽样方法经常用于邮寄调查,因为这种调查的回答率低,并且通过更多的加倍努力(电话或访问)可以从无回答子样本中得到较高的回答百分比。

§ 9.2 为分层的二重抽样

分层抽样是一种应用广泛的抽样方法,但进行分层抽样有一个前提,即需要将总体 N 个单元划分成 L 个互不重叠的层,而且需要知道各层的权重 $W_h = \frac{N_h}{N}$ 。如果事先无法知道总体的层权,可以采用二重抽样方法。

一、符号说明

用下标 h 表示层数, $h = 1, 2, \dots, L$ 。

总体第 h 层的单元数: N_h

总体单元数: $N = \sum_{h=1}^L N_h$

第一重样本第 h 层的单元数: n'_h

第一重样本单元数: $n' = \sum_{h=1}^L n'_h$

第二重样本第 h 层的单元数: n_h

第二重样本单元数: $n = \sum_{h=1}^L n_h$

总体单元第 h 层的权重: $W_h = \frac{N_h}{N}$

第一重样本第 h 层的权重: $w'_h = \frac{n'_h}{n'}$

第二重样本第 h 层的抽样比: $f_{hD} = \frac{n_h}{n'_h}$, $0 < f_{hD} \leq 1$

第二重样本第 h 层 j 单元的观测值: y_{hj} , $j = 1, 2, \dots, n_h$; $h = 1, 2, \dots, L$

第二重样本第 h 层样本单元的平均数: $y_h = \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}$

总体方差: S^2

第 h 层的总体方差: S_h^2

第一重样本第 h 层方差: $s_h'^2$

第二重样本第 h 层方差: $s_h^2 = \frac{1}{n_h - 1} \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2$

二、抽样方法

第一步:利用简单随机抽样,从总体的 N 个单元中随机抽取第一重样本,样本单元数为 n' ;根据已知的分层标志将第一重样本分层,令 $w_h' = \frac{n_h}{n'} (h = 1, 2, \dots, L)$, 则 w_h' 是总体层权 W_h 的无偏估计

第二步:利用分层随机抽样,从第一重样本中抽取出第二重样本,样本单元数为 n ,第 h 层样本单元数为 n_h , $n = \sum_{h=1}^L n_h$ 。

三、估计量及其性质

(一) 均值估计量

采用二重分层抽样,对总体均值 Y 的估计量为:

$$y_{stD} = \sum_{h=1}^L w_h' y_h \quad (9.1)$$

(二) 估计量 y_{stD} 的性质

性质1 估计量 y_{stD} 是 Y 的无偏估计。即

$$E(y_{stD}) = Y \quad (9.2)$$

证明:第二重样本是利用分层随机抽样从第一重样本中抽出的子样本,因此第二重样本第 h 层样本均值 y_h 是第一重样本第 h 层均值 \bar{y}_h' 的无偏估计,即 $E(y_h) = \bar{y}_h'$ 。则在两次抽样下:

$$\begin{aligned} E(y_{stD}) &= E_1[E_2(y_{stD})] = E_1\left[E_2\left(\sum_{h=1}^L w_h' y_h\right)\right] \\ &= E_1\left(\sum_{h=1}^L w_h' \bar{y}_h'\right) = E_1(\bar{y}') = Y \end{aligned}$$

性质2 y_{stD} 的方差为:

$$V(y_{stD}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \sum_{h=1}^L \frac{W_h S_h^2}{n'} \left(\frac{1}{f_{hD}} - 1\right) \quad (9.3)$$

式中, S^2 为总体方差; S_h^2 为第 h 层的总体方差; f_{hD} 为第二重样本第 h 层的抽样比。

性质3 $V(y_{stD})$ 的样本估计量为:

$$v(y_{stD}) = \sum_{h=1}^L \left(\frac{1}{n_h} - \frac{1}{n'_h} \right) w_h'^2 s_h^2 + \left(\frac{1}{n'} - \frac{1}{N} \right) \sum_{h=1}^L w_h' (y_h - y_{stD})^2 \quad (9.4)$$

式中, $v(y_{stD})$ 为 $V(y_{stD})$ 的近似无偏估计; s_h^2 为第二重样本第 h 层方差。

当二重抽样比 $\frac{n_h}{n'}$ 和 $\frac{n'}{N}$ 都可以忽略不计时, (9.4) 式可简化为

$$v(y_{stD}) \approx \sum_{h=1}^L \frac{w_h'^2 s_h^2}{n_h} + \frac{1}{n'} \sum_{h=1}^L w_h' (y_h - y_{stD})^2 \quad (9.5)$$

【例 9.1】某银行要调查其客户的资产情况。已知该银行的客户数为 8 000, 针对客户规模差异较大的特点, 拟采用分层抽样。但由于缺乏现有的分层资料, 决定采用二重分层抽样方法。第一重样本量 $n' = 1\,000$, 根据其自报的资产情况可分为 4 层: 第一层为 300 万元以下; 第二层为 300 万元 ~ 1 000 万元; 第三层为 1 000 万元 ~ 2 000 万元; 第四层为 2 000 万元以上。然后在第一重样本分层的基础上, 在各层分别抽取第二重样本。第二重样本量 $n = \sum_{h=1}^4 n_h = 200$ 。对这 200 个客户进行详细的调查, 取得有关数据整理如表 9.1, 试估计该银行所有客户的资产总额及其抽样标准误差。

表 9.1 某银行客户的样本数据

分 层	第 一 重 样 本	第 二 重 样 本	样本均值(y_h) (百万元)	$\sum y_h^2$	s_h^2
300 万元以下	540	80	2	400	1.01
300 万元 ~ 1 000 万元	320	60	7	3 100	2.71
1 000 万元 ~ 2 000 万元	100	40	15	9 600	15.38
2 000 万元以上	40	20	40	45 120	690.53
合计	1 000	200			

解: 根据表 9.1, 可计算各层的权重:

$$w_1' = 0.54 \quad w_2' = 0.32 \quad w_3' = 0.10 \quad w_4' = 0.04$$

(1) 根据式(9.1), 该银行客户的平均资产额估计为:

$$\begin{aligned} y_{stD} &= \sum_{h=1}^L w_h' y_h \\ &= 0.54 \times 2 + 0.32 \times 7 + 0.10 \times 15 + 0.04 \times 40 \\ &= 6.42 \text{ (百万元)} \end{aligned}$$

该银行共有 8 000 个客户, 故全部客户资产总额为:

$$\hat{Y} = N y_{stD} = 8\,000 \times 6.42 = 51\,360 \text{ (百万元)}$$

(2) 根据式(9.4), y_{sD} 的方差估计为:

$$\begin{aligned}
 v(y_{sD}) &= \sum_{h=1}^L \left(\frac{1}{n_h} - \frac{1}{n'} \right) w_h'^2 s_h^2 + \left(\frac{1}{n'} - \frac{1}{N} \right) \sum_{h=1}^L w_h' (y_h - y_{sD})^2 \\
 \sum_{h=1}^L \left(\frac{1}{n_h} - \frac{1}{n'} \right) w_h'^2 s_h^2 &= \left(\frac{1}{80} - \frac{1}{540} \right) (0.54)^2 (1.01) \\
 &\quad + \left(\frac{1}{60} - \frac{1}{320} \right) (0.32)^2 (2.71) \\
 &\quad + \left(\frac{1}{40} - \frac{1}{100} \right) (0.1)^2 (15.38) \\
 &\quad + \left(\frac{1}{20} - \frac{1}{40} \right) (0.04)^2 (690.53) \\
 &= 0.036\ 822 \\
 \left(\frac{1}{n'} - \frac{1}{N} \right) \sum_{h=1}^L w_h' (y_h - y_{sD})^2 &= \left(\frac{1}{1\ 000} - \frac{1}{8\ 000} \right) [0.54(2 - 6.42)^2 \\
 &\quad + 0.32(7 - 6.42)^2 + 0.1(15 - 6.42)^2 \\
 &\quad + 0.04(40 - 6.42)^2] \\
 &= 0.055\ 239
 \end{aligned}$$

因此, $v(y_{sD}) = 0.036\ 822 + 0.055\ 239 = 0.092\ 061$

该银行客户资产总额的抽样标准误的估计:

$$s(\hat{Y}) = Ns(y_{sD}) = N\sqrt{v(y_{sD})} = 2\ 427.32 \text{ (百万元)}$$

四、二重分层抽样样本量的最优分配

二重分层抽样中有两次抽样,这两次抽样的样本量,即 n' 和 n ,直接影响估计的精度。第一重抽样 n' 越大,对分层信息的了解和估计就越精确,从而可以减少估计量的方差;同样,第二重抽样 n 越大,估计量的方差越小。调查经费是有限的,因此需要在给定费用的条件下,选择 n' 和 n ,使得估计量的方差 $V(y_{sD})$ 最小。

假设第一重抽样的单元平均调查费用为 c_1 (一般情况下,第一重抽样的单元平均调查费用都比较小),第二重抽样第 h 层的单元平均调查费用为 c_{2h} ($h = 1, 2, \dots, L$)。忽略其他费用,则费用函数可以表示为:

$$C_T = c_1 n' + \sum_{h=1}^L c_{2h} n_h \quad (9.6)$$

由于 n_h 是随机变量,所以选择 n' 和 f_{hD} 的期望费用 C_T^* 为:

$$C_T^* = E(C_T) = c_1 n' + n' \sum_{h=1}^L c_{2h} f_{hD} W_h \quad (9.7)$$

根据式(9.3), 总体均值估计量的方差为:

$$V(y_{stD}) = \left(\frac{1}{n'} - \frac{1}{N} \right) S^2 + \sum_{h=1}^L \frac{W_h S_h^2}{n'} \left(\frac{1}{f_{hD}} - 1 \right) \quad (9.8)$$

要在一定的费用约束下令估计方差最小化, 则有

$$\begin{aligned} L = V(y_{stD}) + \lambda (C_T^* - c_1 n' - n' \sum_{h=1}^L c_{2h} f_{hD} W_h) \\ = \left(\frac{1}{n'} - \frac{1}{N} \right) S^2 + \sum_{h=1}^L \frac{W_h S_h^2}{n'} \left(\frac{1}{f_{hD}} - 1 \right) \\ + \lambda (C_T^* - c_1 n' - n' \sum_{h=1}^L c_{2h} f_{hD} W_h) \end{aligned} \quad (9.9)$$

由 $\frac{\partial L}{\partial n'} = 0$ 及 $\frac{\partial L}{\partial f_{hD}} = 0$ 得

$$\begin{cases} f_{hD} = S_h \sqrt{\frac{c_1}{c_{2h} (S^2 - \sum_{h=1}^L W_h S_h^2)}} \\ n' = \frac{C_T^*}{c_1 + \sum_{h=1}^L c_{2h} W_h f_{hD}} \end{cases} \quad (9.10)$$

在实际应用中, 要确定最优的 n' 和 f_{hD} , 需要对总体事先有一定的了解, 例如对 S^2, S_h^2, W_h 有一些粗略的估计。

§ 9.3 为比率估计的二重抽样

第4章介绍了比率估计方法, 通常只要目标变量与辅助变量存在较好的正相关关系, 比率估计的估计精度就优于简单估计。但使用比率估计的前提是已知辅助变量的有关信息。假设研究的变量为 Y (此时 Y 表示目标变量), 辅助变量为 X (此时 X 表示辅助变量), 在估计总体均值 Y 时, 需要辅助变量总体均值 \bar{X} , 才能计算 $\hat{\bar{Y}} = \hat{R}\bar{X}$; 在估计总体总量 Y 时, 需要辅助变量总量 X , 才能估计 $\hat{Y} = \hat{R}X$ 。在实际工作中, 如果辅助变量的信息未知, 可以利用二重抽样进行比率估计。本节以对总体均值 Y 的估计为例进行讨论。

一、二重抽样比率估计的抽样方法

第一步:从总体的 N 个单元中随机抽取第一重样本,样本单元数为 n' ;对于第一重样本,仅观测辅助变量信息,用辅助变量的样本均值 $\bar{x}' = \frac{1}{n'} \sum_{i=1}^{n'} x'_i$ 估计总体均值 X 。

第二步:从第一重样本中随机抽取出第二重样本,样本单元数为 n ;对于第二重样本,观测目标变量与辅助变量,并用获得的 y 和 x ,计算 $\hat{R} = \frac{\bar{y}}{\bar{x}}$,构造比率估计。

二、二重抽样的比率估计及其性质

(一) 二重抽样比率估计

二重抽样对总体均值 Y 的比率估计:

$$y_{RD} = \frac{\bar{y}}{\bar{x}} \bar{x}' \quad (9.11)$$

式中, \bar{y}, \bar{x} 分别为第二重样本目标变量与辅助变量的样本平均数; \bar{x}' 为第一重样本辅助变量的平均数。

(二) 二重抽样比率估计的性质

性质 4 与简单随机抽样下的比率估计一样, y_{RD} 是个有偏估计,其偏倚随着样本量的增大而缩小。当第二重样本的样本量 n 足够大时, $y_{RD} - \frac{\bar{y}}{\bar{x}} \bar{x}'$ 是近似无偏的。即

$$E(\bar{y}_{RD}) \approx Y \quad (9.12)$$

因为在第二重样本的 n 足够大时, $E_2(\hat{R}) \approx R'$, 其中 $\hat{R} = \frac{\bar{y}}{\bar{x}}, R' = \frac{\bar{y}'}{\bar{x}'}$, 所以

$$E(y_{RD}) = E_1[E_2(y_{RD})] = E_1[\bar{x}' E_2(\hat{R})] \approx E_1(\bar{y}') = \bar{Y}$$

因此, y_{RD} 是 \bar{Y} 的近似无偏估计。

性质 5 二重抽样比率估计的方差为:

$$\begin{aligned} V(y_{RD}) &= V_1[E_2(y_{RD})] + E_1[V_2(y_{RD})] \\ &\approx V_1[\bar{y}'] + E_1[(\bar{x}')^2 V_2(\hat{R})] \\ &\approx \left(\frac{1}{n'} - \frac{1}{N}\right) S_{y'}^2 + \left(\frac{1}{n'} - \frac{1}{n}\right) (S_y^2 + R^2 S_x^2 - 2RS_{yx}) \quad (9.13) \end{aligned}$$

通常 $\frac{1}{N}$ 可忽略,因此

$$V(y_{RD}) \approx \frac{1}{n} S_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) (R^2 S_x^2 - 2RS_{yx}) \quad (9.14)$$

式中, S_y^2, S_x^2, S_{yx} 分别为 Y 和 X 的总体方差和总体协方差, $R = \frac{Y}{\bar{X}}$,

性质 6 二重抽样比率估计方差的样本估计:

$$v(y_{RD}) = \frac{1}{n} s_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) (\hat{R}^2 s_x^2 - 2\hat{R}s_{yx}) \quad (9.15)$$

式中,使用目标变量 Y 的第二重样本方差 s_y^2 估计总体方差 S_y^2 ;使用辅助变量 X 的第二重样本方差 s_x^2 估计总体方差 S_x^2 ;使用 Y 和 X 的第二重样本协方差 s_{yx} 估计总体协方差;使用第二重样本比例 $\hat{R} = \frac{y}{x}$ 估计总体比例 $R = \frac{Y}{X}$ 。

【例 9.2】 某住宅小区共有 200 个住户,现欲估计小区住户家庭月平均收入的平均水平。家庭收入的数据不易调查,而家庭支出的资料相对容易获取,而且家庭月平均收入与家庭月平均支出之间强相关,因此拟采用二重抽样比率估计方法。先从住户中随机抽取 100 个住户作为第一重样本,调查家庭月平均支出,结果家庭月平均支出的样本均值为 1 500 元;然后从这 100 个住户中随机抽选 10 户作为第二重样本,调查家庭月平均收入和家庭月平均支出,资料见表 9.2。试估计该小区家庭月平均收入,并计算估计量标准差。

表 9.2 某小区住户家庭收支的样本数据 单位:元

样本住户	家庭月平均支出 (X_i)	家庭月平均收入 (Y_i)
1	1 500	2 000
2	1 200	1 800
3	2 000	2 800
4	1 800	2 500
5	1 300	1 900
6	3 000	5 800
7	800	1 300
8	1 400	2 000
9	1 600	2 300
10	1 100	1 600

解:由题知 $\bar{x} = 1 500$,由表 9.2,计算

$$y = 2 400, \bar{x} = 1 570, \hat{R} = 1.528 7$$

$$s_y^2 = 1\,613\,333, s_x^2 = 371\,222.2, s_{xy} = 747\,777.8$$

根据式(9.11),该小区住户的平均家庭月收入估计为:

$$y_{RD} = \frac{y}{x}x' = 1.528\,7 \times 1\,500 = 2\,293(\text{元})$$

根据式(9.15), y_{RD} 的方差估计为:

$$\begin{aligned} v(y_{RD}) &\approx \frac{1}{n}s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)(\hat{R}^2 s_x^2 - 2\hat{R}s_{xy}) \\ &\approx \frac{1\,613\,333}{10} + \left(\frac{1}{10} - \frac{1}{100}\right)(1.528\,7^2 \times 371\,222.2 \\ &\quad - 2 \times 1.528\,7 \times 747\,777.8) \\ &= 33\,646.89 \end{aligned}$$

\bar{y}_{RD} 的标准差的估计为:

$$s(y_{RD}) = \sqrt{v(y_{RD})} = 183.43(\text{元})$$

三、二重抽样比率估计时样本量的最优分配

在给定的费用条件下,选择第一重样本量 n' 和第二重样本量 $n'f$,其中 f 为抽样比,使得估计量的方差 $V(y_{RD})$ 最小。

假设第一重抽样的单元平均调查费用为 c_1 ,第二重抽样的单元平均调查费用为 c_2 , $h = 1, 2, \dots, L$ 。假设费用函数为:

$$C_T^* = c_1 n' + c_2 n = c_1 n' + c_2 n' f \quad (9.16)$$

根据式(9.14),总体均值估计量的方差为:

$$V(y_{RD}) \approx \frac{1}{n}S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)(R^2 S_x^2 - 2RS_{xy}) \quad (9.17)$$

因此要在一定的费用约束下使估计方差最小化,则有

$$\begin{aligned} L = V(y_{RD}) + \lambda(C_T^* - c_1 n' - c_2 n' f) \\ = \frac{1}{n}S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)(R^2 S_x^2 - 2RS_{xy}) + \lambda(C_T^* - c_1 n' - c_2 n' f) \end{aligned} \quad (9.18)$$

由 $\frac{\partial L}{\partial n'} = 0$ 及 $\frac{\partial L}{\partial f} = 0$ 得

$$\begin{cases} f = \sqrt{\frac{c_1(S_y^2 + R^2 S_x^2 - 2RS_{xy})}{c_2(2RS_{xy} - R^2 S_x^2)}} \\ n' = \frac{C_T^*}{c_1 + c_2 f} \end{cases} \quad (9.19)$$

§ 9.4 为回归估计的二重抽样

与比率估计相似,在辅助信息未知时可以采用二重抽样进行回归估计。以下简单介绍为总体均值 Y 的回归估计进行的二重抽样。

一、二重抽样回归估计的抽样方法

第一步:从总体的 N 个单元中随机抽取第一重样本,样本单元数为 n' ;对于第一重样本,仅观测辅助变量信息 x'_i ,用辅助变量的样本均值 $\bar{x}' = \frac{1}{n'} \sum_{i=1}^{n'} x'_i$ 估计其总体均值 X 。

第二步:从第一重样本中随机抽取出第二重样本,样本单元数为 n ;对于第二重样本,观测目标 y_i 变量与辅助变量 x_i ,并计算 y, x 和回归系数 b ,构造回归估计。

二、二重抽样的回归估计及其性质

(一) 二重抽样回归估计

二重抽样对总体均值 \bar{Y} 的回归估计:

$$y_{lrD} = \bar{y} + b(\bar{x}' - \bar{x}) \quad (9.20)$$

式中, \bar{x}' 和 \bar{x} 分别为第一重样本和第二重样本中辅助变量的平均数; \bar{y} 为根据第二重样本计算的目标变量的样本平均数, b 为根据第二重样本计算的 y_i 对 x_i 的回归系数。

(二) 二重抽样回归估计的性质

性质 7 \bar{y}_{lrD} 是个有偏估计,其偏倚随着样本量的增大而缩小。当第二重样本的样本量 n 足够大时, $y_{lrD} = \bar{y} + b(\bar{x}' - \bar{x})$ 是近似无偏的。即

$$E(y_{lrD}) \approx Y \quad (9.21)$$

$$E(y_{lrD}) = E_1 E_2(y_{lrD}) = E_1 E_2[\bar{y} + b(\bar{x}' - \bar{x})] \approx E_1[\bar{y}'] = Y$$

性质 8 二重抽样回归估计的方差为:

$$V(y_{lrD}) = V_1[E_2(y_{lrD})] + E_1[V_2(\bar{y}_{lrD})] \quad (9.22)$$

式中, $V_2(\bar{y}_{lrD}) \approx \left(\frac{1}{n} - \frac{1}{n'}\right)s_e^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)(1 - \rho^2)s_x^2$, $s_e^2 = (1 - \rho^2)s_y^2$ 为第一重样本残差方差,其均值近似等于总体残差方差 $S_e^2 = (1 - \rho^2)S_y^2$ 。因此

$$\begin{aligned}
 V(y_{lrD}) &\approx V_1(y) + \left(\frac{1}{n} - \frac{1}{n'}\right)E_1(s_e^2) \\
 &\approx \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_y^2(1 - \rho^2) \\
 &\approx \frac{S_y^2}{n} - \left(\frac{1}{n} - \frac{1}{n'}\right)S_y^2\rho^2
 \end{aligned} \tag{9.23}$$

性质 9 二重抽样回归估计方差的样本估计:

$$v(y_{lrD}) = \frac{s_y^2}{n} - \left(\frac{1}{n} - \frac{1}{n'}\right)r^2s_y^2 \tag{9.24}$$

式中是用第二重样本的方差 s_y^2 估计 S_y^2 , 用相关系数 r 估计 ρ 。

【例 9.3】 以例 9.2 的数据, 用二重抽样进行回归估计。试估计该小区家庭月平均收入, 并计算估计量标准差。

解: 由题知 $x' = 1\,500$, 由表 9.2, 计算

$\bar{y} = 2\,400$, $\bar{x} = 1\,570$, 相关系数 $r = 0.966\,26$, 回归系数 $b = 2.014$

$s_y^2 = 1\,613\,333$, $s_x^2 = 371\,222.2$

根据式(9.20), 该小区住户的平均家庭月收入估计为:

$$\begin{aligned}
 y_{lrD} &= \bar{y} + b(x' - \bar{x}) \\
 &= 2\,400 + 2.014 \times (1\,500 - 1\,570) \\
 &= 2\,259.02(\text{元})
 \end{aligned}$$

根据式(9.24), y_{lrD} 的方差估计为:

$$\begin{aligned}
 v(y_{lrD}) &= \frac{s_y^2}{n} - \left(\frac{1}{n} - \frac{1}{n'}\right)r^2s_y^2 \\
 &= \frac{1\,613\,333}{10} - \left(\frac{1}{10} - \frac{1}{100}\right) \times 0.966\,26^2 \times 1\,613\,333 \\
 &= 25\,766.13
 \end{aligned}$$

y_{lrD} 标准差的估计:

$$s(y_{lrD}) = \sqrt{v(y_{lrD})} = 160.52(\text{元})$$

以上例子只是用于说明估计过程。实际应用中, 二重样本容量 n 较大条件下, 才能有效消除用样本回归系数进行回归估计可能产生的偏倚。

三、二重抽样回归估计时样本量的最优分配

在给定的费用条件下, 选择第一重样本量 n' 和第二重样本量 $n''f$, 其中 f 为抽样比, 使得估计量的方差 $V(y_{lrD})$ 最小。

假设第一重抽样的单元平均调查费用为 c_1 , 第二重抽样的单元平均调查费用

为 $c_2, h = 1, 2, \dots, L$ 。假设费用函数为:

$$C_T^* = c_1 n' + c_2 n = c_1 n' + c_2 n f \quad (9.25)$$

根据式(9.23), 总体均值估计量的方差为:

$$V(y_{lrD}) \approx \left(\frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) S_y^2 (1 - \rho^2)$$

要在一定的费用约束下使估计方差最小化, 则有

$$L = V(y_{lrD}) + \lambda (C_T^* - c_1 n' - c_2 n' f) \\ \left(\frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) S_y^2 (1 - \rho^2) + \lambda (C_T^* - c_1 n' - c_2 n' f) \quad (9.26)$$

由 $\frac{\partial L}{\partial n} = 0$ 及 $\frac{\partial L}{\partial f} = 0$ 得

$$\begin{cases} f = \sqrt{\frac{c_1(1-\rho^2)}{c_2\rho^2}} \\ n' = \frac{C_T^*}{c_1 + c_2 f} = \frac{C_T^* \rho}{c_1 \rho + \sqrt{c_1 c_2 (1 - \rho^2)}} \end{cases}$$

小 结

本章介绍了二重抽样的理论及不同目的下二重抽样的估计方法和样本量的分配。二重抽样的主要特点是分两步进行抽样, 每步都抽取一个样本, 而且对每个样本都要获取信息。二重抽样有多种用途, 如有助于筛选主调查对象、节约调查费用、提高调查效率、降低无回答偏倚、研究样本轮换等。

二重分层抽样中有两次抽样, 这两次抽样的样本量, 即 n' 和 n , 直接影响估计的精度。在给定的费用条件下, 选择 n' 和 n 应使得估计量的方差最小。

本章附录 二重抽样公式的证明

1. 分层二重抽样估计的方差(9.3)式的证明。

y_{sD} 的方差为:

$$V(y_{sD}) = \left(\frac{1}{n'} - \frac{1}{N} \right) S^2 + \sum_{h=1}^L \frac{W_h S_h^2}{n'} \left(\frac{1}{f_{hD}} - 1 \right)$$

式中, S^2 为总体方差; S_h^2 为第 h 层的总体方差。

$$\text{证明: } V(y_{stD}) = V_1 \left[E_2 \left(\sum_{h=1}^L w_h y_h \right) \right] + E_1 \left[V_2 \left(\sum_{h=1}^L w_h y_h \right) \right]$$

当 w_h 固定时, $E_2(y_h) = y_h$ 。故有

$$V_1 \left[E_2 \left(\sum_{h=1}^L w_h y_h \right) \right] = V_1 \left(\sum_{h=1}^L w_h y_h \right) = V_1(y') = \left(\frac{1}{n'} - \frac{1}{N} \right) S^2$$

式中, S^2 为总体方差。

当 w_h 固定时, 对第一重样本第 h 层方差 $s_h'^2$, 有 $E_1(s_h'^2) = S_h^2$ 。

$$\begin{aligned} E_1 \left[V_2 \left(\sum_{h=1}^L w_h y_h \right) \right] &= E_1 \left[\sum_{h=1}^L w_h'^2 V_2(y_h) \right] \\ &= E_1 \left[\sum_{h=1}^L w_h'^2 s_h'^2 \left(\frac{1}{n_h} - \frac{1}{n_h'} \right) \right] \\ &= E_1 \left[\sum_{h=1}^L \frac{w_h' s_h'^2}{n'} \left(\frac{1}{f_{hD}} - 1 \right) \right] \\ &= \frac{1}{n'} \sum_{h=1}^L \left(\frac{1}{f_{hD}} - 1 \right) E_1 E_2(w_h' s_h'^2 | w_h' \text{ 固定}) \\ &= \frac{1}{n'} \sum_{h=1}^L \left(\frac{1}{f_{hD}} - 1 \right) E_1(w_h' S_h^2) \\ &= \sum_{h=1}^L \frac{W_h S_h^2}{n'} \left(\frac{1}{f_{hD}} - 1 \right) \end{aligned}$$

$$\text{因此 } V(y_{stD}) = \left(\frac{1}{n'} - \frac{1}{N} \right) S^2 + \sum_{h=1}^L \frac{W_h S_h^2}{n'} \left(\frac{1}{f_{hD}} - 1 \right)$$

2. 二重抽样回归估计的方差(9.23)式的证明。

$$V(y_{lrD}) \approx \frac{S_y^2}{n} - \left(\frac{1}{n} - \frac{1}{n'} \right) S_y^2 \rho^2$$

$$\text{证明: } V(y_{lrD}) = V_1[E_2(\bar{y}_{lrD})] + E_1[V_2(\bar{y}_{lrD})]$$

式中, $V_2(\bar{y}_{lrD}) \approx \left(\frac{1}{n} - \frac{1}{n'} \right) s_e^2 = \left(\frac{1}{n} - \frac{1}{n'} \right) (1 - \rho'^2) s_y^2$; $s_e^2 = (1 - \rho'^2) s_y^2$ 为第一重样本残差方差, 其均值近似等于总体残差方差 $S_e^2 = (1 - \rho^2) S_y^2$ 。因此

$$\begin{aligned} V(\bar{y}_{lrD}) &\approx V_1(\bar{y}') + \left(\frac{1}{n} - \frac{1}{n'} \right) E_1(s_e^2) \\ &\approx \left(\frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) S_y^2 (1 - \rho^2) \\ &\approx \frac{S_y^2}{n} - \left(\frac{1}{n} - \frac{1}{n'} \right) S_y^2 \rho^2 \end{aligned}$$

$$\text{因此 } V(v_{xD}) \approx \frac{S_y^2}{n} \left(\frac{1}{n} - \frac{1}{n'} \right) S_v^2 \rho^2$$

习 题

1. 某县欲调查棉花的播种面积,为及时获取数据,拟采用二重抽样方法。全县共有2 000个村,先抽取500个村作为第一重样本,并根据村的大小进行分层;然后在各层中等比例抽取1/5的村,作为第二重样本,分别调查棉花的种植面积,有关数据如下:

村规模	第一重样本	第二重样本	$\sum v_{hj}$	$\sum y_{hj}^2$
50户以下	85	17	490	15 574
50户 ~ 99户	125	25	1 806	135 164
100户 ~ 199户	140	28	4 423	736 075
200户 ~ 299户	110	22	5 607	1 446 987
300户以上	40	8	4 101	2 205 691

试估计全县棉花的种植面积,并计算估计量标准差。

2. 现有总调查费用3 000元,拟用来做一次比例估计的调查。假设每一个抽样单元的调查费用为10元,拟采用二重分层抽样,第一重样本用于分层,每个抽样单元的分层费用是0.25元。已知总体分为两层,而且两层的权重相等。假如第一层的真实比例为0.2,第二层的真实比例为0.8(假设抽样比 $\frac{n'}{N}$ 和 $\frac{n_h}{N_h}$ 均可忽略不计)。

(1) 试确定二重抽样中最优的 n 和 n' ,以及由此所得的 $V(p_{xD})$;

(2) 试比较二重分层抽样与不分层的简单随机抽样哪个抽样效率高;

(3) 假设每个单元的调查费用为 c_2 ,每个单元的分层费用为 c_1 ,试估计 $\frac{c_2}{c_1}$ 达到多少时二重抽样的费用效率高于简单随机抽样。

3. 某地区预估计牛的年末头数,以上一次的普查数作为辅助变量。但由于行政区划的变动,上次该地区普查的总头数已不能利用,故采用二重抽样的方法,先在全地区1 238个村中抽500个村,得到上期普查数据为平均每村有牛602头,然后又抽取第二重样本为24个村,分别取得上次普查头数和当年的年末头数。其资料如下:

样 本	普查头数	年末头数	样 本	普查头数	年末头数
1	623	654	13	706	707
2	690	696	14	1 795	1 890
3	534	530	15	1 406	1 123
4	293	315	16	118	115
5	69	78	17	330	375
6	842	640	18	218	212
7	475	692	19	160	147
8	371	292	20	210	297
9	161	210	21	262	401
10	298	555	22	204	252
11	2 045	2 110	23	185	199
12	1 069	592	24	574	564

- (1) 试用二重比率估计法估计该地区年末牛的总头数及其估计的标准差;
- (2) 使用二重回归估计法估计该地区年末牛的总头数及其估计的标准差;
- (3) 比较回归估计和比率估计的效率,并做简要分析。

4. 在二重回归抽样中,如果 $\rho = 0.8$,假如由于第一重样本的均值的抽样误差使精确度的损失必须小于 10%,则相对于 n 来说, n' 必须多大?

5. 在二重回归抽样中,假设已知第一重抽样的样本量 $n = 500$,第二重抽样的样本量 $n' = 100$ 。而且对第二重样本,有

$$\sum (y_i - \bar{y})^2 = 17\,283, \sum (x_i - \bar{x})^2 = 3\,248$$

$$\sum (y_i - \bar{y})(x_i - \bar{x}) = 5\,114$$

请计算 Y 的二重回归估计量的标准差。

6. 某小区预调查居民上网情况,为估计居民平均月上网时数,现有两种抽样方案:一种是简单随机抽样,在有限的条件下,只能抽 100 个样本, $v(y_{srs}) = 6.2$;另一种是二重分层抽样,第一重样本用于分层,将居民按月平均收入分为两层,中低收入层($\leq 1\,500$ 元)和高收入层($> 1\,500$ 元),设第一重样本平均分层费用为 c_1 。已知总体数据如下:

层	W_i	S_i^2	S_i	\bar{Y}_i
中低收入层	0.786	312	17.7	19.404
高收入层	0.214	922	30.4	51.626
总体		620	—	26.300

假定总调查费用 C_T 为 100, 第二重样本平均调查费用 $c_{2h} = 1$, 设 $\frac{S^2}{N}$ 可以忽略不计。

(1) 如果 $c_1 = \frac{c_{2h}}{100}$, 试计算二重抽样的样本最优分配方案, 并计算得到的 $v(y_{stD})$;

(2) $\frac{c_1}{c_{2h}}$ 为何值时, 二重抽样的精度高于简单随机抽样。

7. 设总体包含大小相等的 L 个层, 对它采用分层二重抽样, 假设 N 很大, 且第二重抽样的抽样比对各层皆为常数 γ 。试证分层二重抽样估计量 y_{stD} 的方差 $V(y_{stD})$ 满足:

$$nV(y_{stD}) \approx S_h^2 + \frac{n}{n'} \frac{1}{L} \sum_{h=1}^L (Y_h - Y)^2$$

式中, $S_h^2 = \frac{1}{L} \sum_h S_h^2$ 。



第 10 章

复杂样本的方差估计

在前面各章中,我们介绍了几种最基本的抽样方法,并且讨论了在这几种基本抽样方法下比较简单的方差估计问题。但是,实际调查中所面临的情况要复杂得多。首先,实际调查中所使用的抽样方法常常不是简单的一种,而是这些最基本抽样方法的组合,其估计公式比较复杂。其次,在实际调查中抽样的具体实施可能会与最初的抽样设计有一定的差距,因而所得样本是非常复杂的,可称之为复杂样本。这样,按照一般的方法进行复杂样本的方差估计就十分困难。

本章首先对复杂样本调查和复杂样本的方差估计作简单概述,然后分别介绍几种复杂样本的方差估计方法,包括随机组方法、平衡半样本方法、刀切法以及泰勒级数法,最后对这些方法进行比较总结。

§ 10.1 引言

一、复杂样本调查的特点

复杂样本就是从 一个复杂抽样调查所得到的样本。复杂抽样调查主要有以下两个特点。

一是抽样设计复杂。复杂抽样设计经常包括分层、多阶段抽样、不等概率抽样、重复抽样及多个抽样框抽样等内容。

二是调查估计量复杂。复杂调查估计量常常包括比率估计或回归估计等非线性估计量。有时需要对数据进行一些调整,例如无回答的加权调整或插补处理、调整离群值等,这些调整显然会增加调查统计量的复杂程度。

此外,复杂样本调查还可能涉及多变量问题,包含数十个甚至数百个感兴趣的指标,并且调查的规模大、范围广。

二、复杂样本方差估计考虑的因素

抽样调查工作者一般面临着两个必须解决的问题:一是构造一个合适的统计量,对感兴趣的总体指标(参数)进行估计;二是对每个估计量的精度进行度量。对精度的度量最常用的是调查估计量的方差。一般地讲,方差是未知的,只能从调查数据本身来估计。调查统计量的方差是由统计量本身的形式和抽样方案设计的性质这两方面决定的。

那么,对于一个复杂样本,怎样为调查估计量选择一个合适的方差估计呢?这是一个相当困难的问题,它涉及方差估计的精度、费用和时间以及操作的简便性等因素,调查工作者要综合考虑这些因素并做出一定的权衡。

(一) 精度

方差估计量的精度可用许多方法衡量,一个重要的度量是方差估计量的均方误差(MSE)。按照这个标准,具有最小均方误差的估计量最好。由于方差估计值经常要用来构造主要调查参数的区间估计,所以精度的第一个标准必须与得到的区间的质量有关,最好是给出最优区间估计的方差估计量。然而,这些标准之间可能存在着矛盾。此外,调查可能包括多变量、时间序列和对调查数据的其他统计分析,这时应选用对要进行的分析有最佳统计性质的方差估计量。一般地,由于对相同数据的不同分析要用不同的方差估计量,所以必须采用折衷的办法。

(二) 费用和时间

虽然精度问题对选择方差估计量有决定性的影响,但是费用和时间因素也起重要作用,对复杂抽样调查更是如此。这类调查可能包括数十张统计表,每张表可能有上百个或更多个数据,也包括回归系数、相关系数等估计量。对每个调查统计量都计算高精度的方差估计值,其费用可能相当惊人,甚至可能超过调查总预算的费用。这时,可能更需要节约费用的方差估计方法,即使这些方法在精度方面可能要损失一些时间也是一个在实际应用中需要考虑的重要因素,因为复杂抽样调查一般都有相当严格的完成时间和发布期限。

(三) 操作的简便性

首先,大多数复杂抽样调查涉及很多变量和统计量及其相应的方差估计值、理论上对每个调查统计量应该使用不同的方差估计量,至少对不同类型的统计量使用不同的方差估计量。然而在许多实际调查中,由于调查经费、专业人员、时间和计算机等资源的稀缺,往往采用折衷的办法,选择一个可能对任何一个统计量都不是最好的,但对所有或者至少对最重要的那些调查统计量来说是精度方面损失最少的方差估计量。其次,在没有合适的软件进行数据处理和方差估计时,需要编制专用的计算机程序。如果编程人员不能正确地编制恰当的计算机程序,特别强调精确的方差估计方案就没有意义。最后,通常来说,使用较简单的估计方法将有利于调查主办者和调查数据的其他使用者的理解,从而可以更好地达到调查目的。

三、典型方法概述

复杂样本的方差估计方法可以分为两类,即重抽样方法和线性化方法。这些方差估计方法所得到的估计量不一定是无偏的,但是有很大的灵活性,能够充分适应复杂抽样调查的大多数特性。本章主要介绍以下几种方差估计方法。

随机组方法是发展最早的一种方差估计方法,它是一种重抽样方法,其实质是按一定的抽样方案从总体中抽取若干组样本,对于每一组样本都建立有关参数的估计量。这些估计量之间的离散程度,即样本方差可用于计算全样本估计量的方差。

平衡半样本方法也是一种重抽样方法,它将各层中的随机组数减为两个,以提高方差估计的效率,但它与随机组方法有所区别。

刀切法建立在再抽样理论上,利用再抽样技巧将原来的总体进行复制,在复制的总体中可以使用原来的抽样办法再复制抽样样本,并构造同样结构的有关参数的统计量。由于复制的总体及统计量是原有总体及统计量的一个缩影,而在复制的模型中,包括统计量的均值、方差等特性在内的指标几乎均可以通过计算得到。

泰勒级数法属于线性化的方差估计方法,其实质是将非线性估计线性化。在抽样调查中人们会遇到一些非线性估计量,如比率估计、回归系数估计等,利用泰勒级数展开可以用线性估计去逼近非线性估计量,从而得到非线性方差估计量的近似估计。不过,泰勒级数法在数学运算上相当复杂。

§ 10.2 随机组方法

方差估计的随机组方法(random group method), 是使用相同的抽样方案从总体中抽取两个或两个以上的样本, 对每一个样本分别构造所感兴趣的总体参数的估计量, 对所有样本的组合再构造一个估计量, 然后利用这些估计量的样本方差或这些估计量与基于全样本的估计量之间的离散程度计算基于全样本估计量的方差。

随机组方法的实质是将抽取的样本分成若干组, 每组子样本作为原始样本的复制, 再利用各子样本估计量之间的离散程度构造方差估计量。由于每个随机组都是原样本或全样本的一个子样本, 且其在整个样本中是交叉散布的, 所以这种方法也称为交叉子样本(interpenetrating subsamples) 方法。该方法是由印度统计学家马哈拉诺比斯(Mahalanobis) 提出来的。

随机组方法有两种基本形式: 一种是随机组之间相互独立; 另一种是随机组之间具有某种相关性。下面分别予以介绍。

一、独立随机组

(一) 随机组的形成

如果每次抽取的样本都被放回, 则所得的随机组为独立的, 具体的抽样过程如下。

1. 按某种抽样方式从总体中抽取样本 S_1 (抽样设计本身没有限制, 可以包括多重抽样框, 多阶段、固定的或随机的样本量, 可以是分层抽样、多阶抽样、多重抽样、放回或不放回抽样)。

2. 在抽取了第一个样本 S_1 后, 将其放回总体, 然后按与之相同的抽样方式抽取样本 S_2 。

3. 重复上述过程, 直至获得 k 个样本 $S_1, S_2, \dots, S_k (k > 2)$ 。我们称这 k 个样本为随机组。

(二) 随机组估计量

对每个随机组, 构造参数 θ 的一个估计量, 分别记为 $\hat{\theta}_\alpha (\alpha = 1, 2, \dots, k)$ 。这样, 方差随机组估计量具有下述性质。

性质 1 设 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ 是相互独立的随机变量, 并且具有相同的期望

$E(\hat{\theta}_a) = \mu$, 定义

$$\hat{\theta} = \frac{1}{k} \sum_a \hat{\theta}_a \quad (10.1)$$

则 $E(\hat{\theta}) = \mu$, $\hat{\theta}$ 的方差 $V(\hat{\theta})$ 的一个无偏估计是

$$v(\hat{\theta}) = \frac{1}{k(k-1)} \left[\sum_{a=1}^k (\hat{\theta}_a - \hat{\theta})^2 \right] \quad (10.2)$$

在实际抽样调查中, 一般使用无偏或近似无偏的估计, 尤其当 $\hat{\theta}_a$ 和 θ 线性时, 期望值 μ 与研究参数 θ 常常是一致或相近似的, 即

$$E(\hat{\theta}) = \mu = \theta \quad (10.3)$$

性质 2 设 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ 服从 $N(\theta, \sigma^2)$ 的独立同分布随机变量, 则

(1) 统计量 $z = \frac{(\hat{\theta} - \theta)}{\sqrt{\frac{\sigma^2}{k}}}$ 服从标准正态分布 $N(0, 1)$;

(2) 统计量 $t = \frac{(\hat{\theta} - \theta)}{\sqrt{v(\hat{\theta})}}$ 服从自由度为 $k - 1$ 的 t 分布。

这样, 如果 $\hat{\theta}_a$ 的方差已知, 或随机组数 k 很大, 则 θ 的 $1 - \alpha$ 的置信区间为:

$$\left(\hat{\theta} - \mu_{\frac{\alpha}{2}} \sqrt{v(\hat{\theta})}, \hat{\theta} + \mu_{\frac{\alpha}{2}} \sqrt{v(\hat{\theta})} \right)$$

如果 $\hat{\theta}_a$ 的方差未知, 或随机组数 k 较小, 则 θ 的 $1 - \alpha$ 的置信区间为:

$$\left(\hat{\theta} - t_{k-1, \frac{\alpha}{2}} \sqrt{v(\hat{\theta})}, \hat{\theta} + t_{k-1, \frac{\alpha}{2}} \sqrt{v(\hat{\theta})} \right)$$

在实际抽样调查中, 当 $\hat{\theta}_a$ 和 θ 非线性时, $\hat{\theta}$ 的期望值 μ 与研究参数 θ 之间存在偏倚 $\mu - \theta \neq 0$ 。但在现代复杂抽样调查中都使用大样本, 这种偏倚通常并不重要。此外, 现实 $\hat{\theta}_a$ 的正态性假定往往不能满足, 但大样本下 $\hat{\theta}_a$ 具有渐进正态分布。

(三) 估计量 $\hat{\theta}$ 的方差估计

一般而言, 估计量 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ 常常是用相同的方式构造出来的。自然地, 我们可以基于 k 个随机组的联合样本, 用构造 $\hat{\theta}_a$ 同样的方式构造 θ 的估计量 $\hat{\theta}$ 而不是简

单地将 $\hat{\theta}_a$ 进行算术平均。显然,对于线性估计量, $\hat{\theta}$ 与 $\hat{\bar{\theta}}$ 是一致的;但对于非线性估计量, $\hat{\theta}$ 与 $\hat{\bar{\theta}}$ 并不相同。

对于估计量 $\hat{\theta}$ 的方差,以下两个估计量都可使用:

$$v_1(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{a=1}^k (\hat{\theta}_a - \hat{\theta})^2 \quad (10.4)$$

$$v_2(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{a=1}^k (\hat{\theta}_a - \hat{\bar{\theta}})^2 \quad (10.5)$$

显然, $v_1(\hat{\theta}) = v_2(\hat{\bar{\theta}})$,也就是说这个估计量不仅可以用于估计 $\hat{\theta}$ 的方差,也可以用于估计 $\hat{\bar{\theta}}$ 的方差。一般来说 $V(\hat{\theta})$ 与 $V(\hat{\bar{\theta}})$ 相差不大,因此 $v_1(\hat{\theta})$ 是 $V(\hat{\theta})$ 的一个合理的估计。

对于线性估计量, $\hat{\theta} = \hat{\bar{\theta}}$,因此 $v_1(\hat{\theta}) = v_2(\hat{\theta})$ 。

对于非线性估计量, $\hat{\theta}$ 与 $\hat{\bar{\theta}}$ 并不相同。

$$\sum_{a=1}^k (\hat{\theta}_a - \hat{\theta})^2 = \sum_{a=1}^k (\hat{\theta}_a - \hat{\bar{\theta}})^2 + k(\hat{\bar{\theta}} - \hat{\theta})^2$$

故有 $v_1(\hat{\theta}) \leq v_2(\hat{\theta})$

上面的关系并不意味着 $v_1(\hat{\theta})$ 就比 $v_2(\hat{\theta})$ 好。因为我们的目的是要估计 $\hat{\theta}$ 的方差而不是要给出一个最小的估计。事实上,出于稳妥的考虑,人们常常更愿意取保守的 $v_2(\hat{\theta})$ 作为 $V(\hat{\theta})$ 的估计。由于在许多复杂样本的调查中, $E(\hat{\theta} - \hat{\bar{\theta}})^2$ 的值经常不是很大,因而 v_1 与 v_2 之间的差别其实并不很大。至于 v_1 与 v_2 到底哪个更好,目前尚无定论,这是一个有待解决的问题。现实中 v_1 与 v_2 几乎是一样的,如果两者出现显著差异,很可能是计算错误或小样本引起的偏倚。

【例 10.1】拒答率调查

为研究被调查者拒答情况,实施一项调查。抽样方式为两阶段抽样,第一阶段采用分层随机抽样从各城区中随机抽取居委会,假设各层权重相同。第二阶段从被抽中的居委会随机抽取住户。抽出样本 S_1 后将其放回总体,然后再按相同的抽样方式抽取出样本 S_2 ,两样本的拒答情况统计结果见表 10.1。试利用随机组方法估计拒答率 \hat{R} 的方差。

表 10.1

样本的拒答情况

城区	样本 S_1		样本 S_2	
	拒答户数(y_1)	合格调查户数(x_1)	拒答户数(y_2)	合格调查户数(x_2)
1	41	150	37	149
2	40	149	30	148
3	38	145	38	150
总计	119	444	105	447

解:根据样本 S_1 数据,估计拒答率 $\hat{R}_1 = \frac{\hat{Y}_1}{\hat{X}_1} = \frac{\sum_{i=1}^3 y_{1i}}{\sum_{i=1}^3 x_{1i}} = \frac{119}{444} = 0.268\ 018$

根据样本 S_2 数据,估计拒答率 $\hat{R}_2 = \frac{\hat{Y}_2}{\hat{X}_2} = \frac{\sum_{i=1}^3 y_{2i}}{\sum_{i=1}^3 x_{2i}} = \frac{105}{447} = 0.234\ 899$

然后再基于样本 S_1 和样本 S_2 的联合数据,估计拒答率:

$$\hat{R} = \frac{\hat{Y}_1 + \hat{Y}_2}{\hat{X}_1 + \hat{X}_2} = \frac{224}{891} = 0.251\ 402\ 92$$

$$\hat{R} = \frac{1}{2}(\hat{R}_1 + \hat{R}_2) = 0.251\ 458\ 67$$

因为比是一个非线性统计量,所以我们用两种方法估计拒答率 \hat{R} 的方差:

$$v_1(\hat{R}) = \frac{1}{k(k-1)} \sum_{a=1}^k (\hat{R}_a - \hat{R})^2 = \frac{1}{2(2-1)} \sum_{a=1}^2 (\hat{R}_a - \hat{R})^2 = 0.000\ 274\ 21$$

$$v_2(\hat{R}) = \frac{1}{k(k-1)} \sum_{a=1}^k (\hat{R}_a - \hat{R})^2 = \frac{1}{2(2-1)} \sum_{a=1}^2 (\hat{R}_a - \hat{R})^2 = 0.000\ 274\ 22$$

很明显,对这些数据来说, \hat{R} 和 \hat{R} 以及 $v_1(\hat{R})$ 和 $v_2(\hat{R})$ 之间的差别很小。

【例 10.2】 AAA 汽车旅馆调查^①

我们现在用美国汽车协会(American Automobile Association)对其会员汽车旅

^① 此例取自戴明(Deming, 1960),转引自 K. M. Wolter:《方差估计引论》,32 页,北京,中国统计出版社,1998。

馆经营者的实际调查来说明独立随机组的使用。该调查的目的是确定这些经营者是否喜欢建立一个预订系统,使汽车驾驶者可以提前预订房间。

调查框是 AAA 中心办公室里的卡片文件,包括 172 个文件抽屉,每个抽屉有 64 张卡片,每张卡片代表一个基本单元(可能是合同汽车旅馆、饭店、空白卡片等),抽样单元是卡片。

调查的抽样设计如下:

1. 事先已知总体中约有 5 000 个合同汽车旅馆,并准备抽取约 700 个单元做总样本,这样,总抽样比约为 $\frac{700}{5\,000}$,即约 7 个里面抽 1 个,所以,每一个抽屉都各增加 6 张空白卡片,这样每个抽屉都有 70 张卡片。

2. 从每一个抽屉中随机抽取一张卡片,组成一个 172 张卡片的样本。抽样在不同的抽屉中是相互独立的。

3. 按照第 2 步的方法有放回地再抽取 9 个样本,这样,由这 10 个样本(或随机组)得出的估计量可以认为是相互独立的。

4. 结果有 854 个汽车旅馆被抽入总样本,向每一个单元寄一张调查表。其他 866 个单元不是合同汽车旅馆,不属于被调查总体。虽然使用有放回的抽样方法抽取随机组,但没有单元被重复抽中。

5. 10 天后,对无回答的单元第二次寄调查表,再过一星期第三次寄调查表。如果 24 天后仍无返回调查表,就被认为是无回答者。

6. 将无回答者按随机组的数字顺序排列,并从每 3 个相邻组中随机抽选一个,对抽中的无回答单元进行面访。在这种抽样中,前一随机组中最后面的无回答者放到下一随机组的前面,这样做有点破坏随机组估计量独立性的条件。然而,在本例中,这一点被忽略。

表 10.2 给出了 24 天后关于问题“人们经常向你预订吗”的结果,表 10.3 给出了无回答子样本对此问题的回答。

表 10.2 24 天后对问题“人们经常向你预定吗”的各类回答结果

随机组	经 常	很 少	没 有	不明确回答	未回答	合 计
1	16	40	17	2	19	94
2	20	30	17	3	15	85
3	18	35	16	1	15	85
4	17	31	14	2	16	80
5	14	32	15	3	18	82
6	15	32	12	4	16	79

续前表

随机组	经 常	很 少	没 有	不明确回答	未回答	合 计
7	19	30	17	3	17	86
8	13	37	11	3	18	82
9	19	39	19	2	14	93
10	17	39	15	2	15	88
合 计	168	345	153	25	163	854

表 10.3 对无回答子样本访问的结果

随机组	经 常	很 少	没 有	暂时关闭(放假、生病等)	合 计
1	1	2	2	1	6
2	1	2	1	1	5
3	2	2	0	1	5
4	2	1	2	0	5
5	1	3	1	2	7
6	2	2	0	1	5
7	1	3	1	1	6
8	1	2	1	2	6
9	2	2	1	0	5
10	1	2	0	2	5
合 计	14	21	9	11	55

给定的样本单元属于任何一个随机组的概率是 $\frac{1}{70}$,属于无回答者子样本的条件概率是 $\frac{1}{3}$ 。这样,来自第一个随机组的合同汽车旅馆的总数估计是:

$$\begin{aligned}\hat{X}_1 &= 70 \sum_{i=1}^{172} X_{1i} \\ &= 70 \times 94 \\ &= 6\,580\end{aligned}$$

式中, $X_{1i} = \begin{cases} 1, & \text{第 } i \text{ 个随机组中的第 } i \text{ 个抽中单元是合同旅馆} \\ 0, & \text{其他} \end{cases}$

所有随机组的总数估计是:

$$\hat{X} = \sum_{a=1}^J \frac{\hat{X}_a}{10} = 5\,978$$

因为估计量是线性的,所以 \hat{X} 和 $\hat{\bar{X}}$ 是相同的,相应的方差估计是:

$$v(\hat{X}) = \frac{1}{10 \times 9} \sum_{a=1}^{10} (\hat{X}_a - \hat{X})^2 = 12\,652.9$$

表 10.4 给出了关于问题“人们经常向你预订吗”的每一分类的总数估计。

表 10.4 总数的估计

随机组	经 常	很 少	没 有	不明确回答	暂时关闭
1	1 330	3 220	1 610	140	210
2	1 610	2 520	1 400	210	210
3	1 680	2 870	1 120	70	210
4	1 610	2 380	1 400	140	0
5	1 190	2 870	1 260	210	420
6	1 470	2 660	840	280	210
7	1 540	2 730	1 400	210	210
8	1 120	3 010	980	210	420
9	1 750	3 150	1 540	140	0
10	1 400	3 150	1 050	140	420
母样本	1 470	2 856	1 260	175	231

例如,第一个随机组中回答“经常”的汽车旅馆总数的估计值是:

$$\begin{aligned}\hat{Y}_1 &= 70 \times \left(\sum_{i \in r_1} Y_{1i} + 3 \sum_{i \in nr_1} Y_{1i} \right) \\ &= 70 \times (16 + 3 \times 1) \\ &= 1\,330\end{aligned}$$

式中, $\sum_{i \in r_1}$ 和 $\sum_{i \in nr_1}$ 分别为对第一个随机组中回答者和无回答者子样本的求和。

$$Y_{1i} = \begin{cases} 1, & \text{第一个随机组中的第 } i \text{ 个抽中单元是合同旅馆并回答“经常”} \\ 0, & \text{其他} \end{cases}$$

各种非线性统计量也可用这些数据来处理。关于第一个随机组中很少 + 没有 / 经常 + 很少 + 没有 的比的估计值是:

$$\hat{R}_1 = \frac{3\,220 + 1\,610}{1\,300 + 3\,220 + 1\,610} = 0.784$$

所有随机组的这个比的估计值是:

$$\hat{R} = \frac{\sum_{a=1}^{10} \hat{R}_a}{10} = 0.737$$

相应的方差估计值是:

$$v(\hat{R}) = \frac{1}{10 \times 9} \sum_{a=1}^{10} (\hat{R}_a - \hat{R})^2 = 0.000\ 113\ 9$$

因为比是一个非线性统计量,所以我们可以使用另一估计值

$$\hat{R} = \frac{2\ 856 + 1\ 260}{1\ 470 + 2\ 856 + 1\ 260} = 0.737$$

$\text{Var}(\hat{R})$ 的两个随机组估计值是:

$$v_1(\hat{R}) = v(\hat{R}) = 0.000\ 113\ 9$$

$$v_2(\hat{R}) = \frac{1}{10 \times 9} \sum_{a=1}^{10} (\hat{R}_a - \hat{R})^2 = 0.000\ 113\ 9$$

二、非独立随机组

在实际应用中,往往很难实现一系列的有放回独立抽样,而经常是采用不放回抽样方法一次性抽取一个整体调查样本,然后将该样本随机地分为 k 个随机组。这种随机组估计量 $\hat{\theta}_g$ 之间就不再独立了。

(一) 随机组的形成

为了保证方差的随机组估计量具有较好的统计性质,随机组的划分必须遵循以下基本原则,即每个随机组本质上具有与原始样本相同的抽样设计。例如,对于一个系统样本,如果从该样本中再以同样的系统抽样法抽取出一个子系统样本,则该子样本可看成是用与原始样本相同的抽样方法抽取出来的,这样的子样本就可作为一个随机组。

非独立随机组的具体形成一般要遵循以下原则。

1. 如果原始样本是用不放回的简单随机抽样或不放回的 π PS 抽样方式抽取的,则随机组可通过随机地划分原始样本得到。具体步骤如下:

- (1) 从原始样本中简单随机地抽取 $m = \frac{n}{k}$ 个单元,形成第一个随机组;
- (2) 再从剩下的 $n - m$ 个单元中简单随机地抽取 m 个单元作为第二个随机组;
- (3) 依此类推,即可得 k 个随机组;
- (4) 如果 $\frac{n}{k}$ 不是整数,不妨设 $n = mk + q$ ($0 < q < k$),那么剩下的 q 个单元要么弃之不用,要么将它们逐一加到前面 q 个随机组中。

2. 如果原始样本是用等概率或不等概率系统抽样方式抽取的,则可通过对原始样本采用系统抽样形成随机组。具体步骤如下:

- (1) 从整数 1 到 k 中随机抽取一个整数, 记为 α^* ;
- (2) 原始样本中第一个单元进入第 α^* 随机组;
- (3) 第二个单元进入第 $\alpha^* + 1$ 随机组;
- (4) 依此类推, 直到取完 k 个随机组。

3. 对于多阶抽样, 将来自同一初级抽样单元 (PSU) 的所有基本样本单元的集合作为一个整体, 称为末级群。随机组是通过将所有末级群分成 k 组而得到的, 具体的划分方法根据第一阶抽样方法而定。如果第一阶抽样是不放回的简单随机抽样或 π PS 抽样, 则使用原则 1; 如果第一阶抽样是系统抽样, 则使用原则 2。

4. 对于分层抽样, 如果希望估计层内方差, 那么在该层内根据所采用的抽样方法而使用原则 1、原则 2 或原则 3; 如果希望估计总体方差, 那么每个随机组本身必须是一个分层样本, 此时应将从每一层中抽得的样本划分成 k 组, 然后在各层中任意取一个随机组, 从而形成原始样本的一个随机组。

5. 如果采用的是二重抽样, 则应将第一重样本按原则 1 或原则 2 划分成 k 个随机组; 而第二重样本则被相应地分成随机组, 即第二重样本单元所在的随机组完全由第一次划分时决定。这种划分的前提当然是第一、二重样本均已被抽取出来。在实际应用中, 有时随机组的划分是在第一重样本抽出后第二重样本抽出前进行的, 这时是将第一重样本按原则 1 或原则 2 分成 k 个随机组, 再从每个随机组中独立地抽取 $m = \frac{n}{k}$ 个单元一起组成第二重样本。

(二) 非独立随机组的估计

对于非独立随机组, 估计方法与独立随机组的情形类似。仍以 $\hat{\theta}$ 表示根据原始样本构造的估计量, 以 $\hat{\theta}_\alpha$ 表示由第 α 个随机组构造的估计量, 则

$$\hat{\theta} = \frac{1}{k} \sum_{\alpha=1}^k \hat{\theta}_\alpha \quad (10.6)$$

$\hat{\theta}$ 的方差 $V(\hat{\theta})$ 的随机组估计为:

$$v(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 \quad (10.7)$$

与独立随机组相同, $\hat{\theta}$ 的方差 $V(\hat{\theta})$ 也可以通过两式估计:

$$v_1(\hat{\theta}) = v(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 \quad (10.8)$$

$$v_2(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 \quad (10.9)$$

出于同样的理由,为保险起见,我们取 $v_2(\hat{\theta})$ 作为 $V(\hat{\theta})$ 的估计。

(三) 非独立随机组估计的性质

性质 3 设 $E(\hat{\theta}_a) = \mu_a, \mu_a$ 不一定等于 μ , 有

$$E(\hat{\theta}) = \frac{1}{k} \sum_{a=1}^k \mu_a = \mu \quad (10.10)$$

且 $V(\hat{\theta})$ 的随机组估计量的期望

$$E[v(\hat{\theta})] = V(\hat{\theta}) + \frac{1}{k(k-1)} \sum_{a=1}^k (\mu_a - \mu)^2 - \frac{2}{k(k-1)} \sum_{a=1}^k \sum_{\beta > a}^k \text{Cov}(\hat{\theta}_a, \hat{\theta}_\beta) \quad (10.11)$$

性质 3 显示,由于非独立性, $v(\hat{\theta})$ 不再是 $V(\hat{\theta})$ 的无偏估计。但若总体较大,抽样比又较小,则 $\frac{2}{k(k-1)} \sum_{a=1}^k \sum_{\beta > a}^k \text{Cov}(\hat{\theta}_a, \hat{\theta}_\beta)$ 常常相对较小且为负值;而当 $\mu_a \approx \bar{\mu}$ 时, $\frac{1}{k(k-1)} \sum_{a=1}^k (\mu_a - \mu)^2$ 也较小。因此,在许多大规模的抽样调查中, $v(\hat{\theta})$ 的偏倚通常不会很大。

三、随机组数 k 的选择

在设计阶段,什么样的调查设计和多大的样本量才能保证调查估计量 θ 的估计精度?要解决这个问题必须先了解 θ 的方差。此外,还应注意 θ 的方差估计的稳定性。在随机组方差估计中,随机组数的选择会影响方差估计的精度。

确定随机组估计量稳定性的一般准则是变异系数准则(CV 准则):

$$CV[v(\hat{\theta})] = \frac{|V[v(\hat{\theta})]|^{\frac{1}{2}}}{V(\hat{\theta})}$$

性质 4 假设 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ 为独立同分布变量, 而 $v(\hat{\theta}) = \frac{1}{k(k-1)} \left[\sum_{a=1}^k (\hat{\theta}_a - \hat{\theta})^2 \right]$, 则 $v(\hat{\theta})$ 的变异系数为:

$$CV[v(\hat{\theta})] = \left[\frac{\beta_4(\hat{\theta}_1) - \frac{k-3}{k-1}}{k} \right]^{\frac{1}{2}} \quad (10.12)$$

$$\text{式中, } \beta_4(\hat{\theta}_1) = \frac{E[(\hat{\theta}_1 - \mu)^4]}{E[(\hat{\theta}_1 - \mu)^2]^2}, \mu = E(\hat{\theta}_1)。$$

由性质4可见,独立随机组方差估计的CV与 $\hat{\theta}_a$ 的分布和随机组组数 k 这两个因素密切相关。峰度 $\beta_4(\hat{\theta}_1)$ 越大,方差估计精度越差。组数 k 越小,方差估计精度越差。而且当 k 较大时, CV^2 近似反比于随机组组数 k :

$$CV^2[v(\hat{\theta})] \approx \frac{\beta_4(\hat{\theta}_1)}{k} \quad (10.13)$$

从方差估计的精度角度出发,显然随机组组数越大越好,但随机组组数 k 的选择还要受成本的约束,最优的随机组组数 k 应该从精度和成本两方面进行权衡。如果调查的目的只是为了得到某总体指标的大致结果,成本因素比精度因素重要,则随机组组数 k 可以小一些;如果要依靠调查结果制定重要决策,精度要求较高,建议采用较大的随机组组数 k 。

§ 10.3 平衡半样本方法

实际分层抽样调查中,出于效率的考虑,每层经常只抽2个单元。在这种情况下,只有2个独立随机组可用于方差估计,使得方差的估计值起伏较大。本节介绍的平衡半样本方法(balanced half-sample method)可以解决这个问题。20世纪50年代末美国普查局的W.N.赫维茨和M.格尼等人提出了方差估计的半样本法的基本思想,后来麦卡锡进一步提出平衡半样本法。

一、半样本

假设对总体 $N = \sum_{h=1}^L N_h$,采用分层随机抽样,每层有放回地简单随机抽取2个单元,设 y_{h1} 和 y_{h2} 是第 h 层的样本观测值($h = 1, 2, \dots, L$),则总体均值 Y 的无偏估计为:

$$y_{st} = \sum_{h=1}^L W_h y_h \quad (10.14)$$

式中, $W_h = \frac{N_h}{N}$ 为层权; $y_h = \frac{y_{h1} + y_{h2}}{2}$ 。

y_{π} 的方差 $V(y_{\pi})$ 的标准估计量为:

$$v(y_{\pi}) = \frac{1}{2} \sum_{h=1}^L W_h^2 d_h^2 = \frac{1}{4} \sum_{h=1}^L W_h^2 d_h^2 \quad (10.15)$$

式中, $d_h = (y_{h1} - y_{h2})$ 。

使用随机组方法, 因每层只抽 2 个单元, 所以只有 2 个独立的随机组 $(y_{11}, y_{21}, \dots, y_{l1})$ 和 $(y_{12}, y_{22}, \dots, y_{l2})$ 。此时 $V(y_{\pi})$ 的随机组估计为:

$$v_{RG}(y_{\pi}) = \frac{1}{2(2-1)} \sum_{a=1}^2 (y_{\pi,a} - y_{\pi})^2 = \frac{1}{4} (y_{\pi,1} - y_{\pi,2})^2 \quad (10.16)$$

式中, $y_{\pi,1} = \sum_{h=1}^L W_h y_{h1}$; $y_{\pi,2} = \sum_{h=1}^L W_h y_{h2}$; $y_{\pi} = \frac{1}{2} (y_{\pi,1} + y_{\pi,2})$ 。

这个估计量计算简单, 但由于仅有一个自由度, 其稳定性比标准估计量 $v(y_{\pi})$ 差。为了既保留随机组估计 $v_{RG}(y_{\pi})$ 的简单性, 又能保持标准估计量 $v(y_{\pi})$ 的稳定性, 我们引入半样本方法, 即从每层抽取一个单元形成半样本, 总共可能出现 2^L 个半样本。由于不同的半样本中包含某些共同的单元, 所以半样本之间是彼此相关的。在这一点上, 半样本方法与随机组方法存在本质上的不同。

二、半样本估计量

假定一个半样本是从每层中抽取一个单元组成, 显然, 对一个给定的样本有 2^L 个这样的半样本。 Y 的基于第 α 个半样本的估计量为:

$$y_{\pi,\alpha} = \sum_{h=1}^L W_h (\delta_{h1\alpha} y_{h1} + \delta_{h2\alpha} y_{h2}) \quad (10.17)$$

式中, $\delta_{h1\alpha} = \begin{cases} 1, & \text{第 } h \text{ 层中第一个单元被选入第 } \alpha \text{ 个半样本} \\ 0, & \text{其他} \end{cases}$
 $\delta_{h2\alpha} = 1 - \delta_{h1\alpha}$

性质 5 所有 2^L 个这样的估计量的平均值恰好是 y_{π} , 即

$$\frac{1}{2^L} \sum_{\alpha=1}^{2^L} y_{\pi,\alpha} = y_{\pi} \quad (10.18)$$

证明: 因为样本中的每个单元都会在一半即 2^{L-1} 个半样本中出现, 即

$$\sum_{\alpha=1}^{2^L} \delta_{h1\alpha} = 2^{L-1}$$

因而 $\frac{1}{2^L} \sum_{\alpha=1}^{2^L} y_{\pi,\alpha} = \sum_{h=1}^L W_h (y_{h1} + y_{h2}) \left(\frac{2^{L-1}}{2^L} \right) = y_{\pi}$

我们将利用 $y_{\pi,\alpha}$ 之间的差异来构造方差估计量。定义

$$\delta_h^{(\alpha)} = \begin{cases} 2\delta_{h1\alpha} - 1 & \text{第 } h \text{ 层中第一个单元被选入第 } \alpha \text{ 个半样本} \\ 1 & \text{第 } h \text{ 层中第二个单元被选入第 } \alpha \text{ 个半样本} \end{cases}$$

则

$$y_{\alpha, a} - y_{\alpha} = \frac{1}{2} \sum_{h=1}^L W_h \delta_h^{(\alpha)} d_h$$

由于 $\delta_h^{(\alpha)2} = 1$, 于是

$$\begin{aligned} (y_{\alpha, a} - y_{\alpha})^2 &= \frac{1}{4} \sum_{h=1}^L W_h^2 d_h^2 + \frac{1}{2} \sum_{h=1}^L \sum_{h'=1}^L \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} W_h W_{h'} d_h d_{h'} \\ &= v(y_{\alpha}) + \frac{1}{2} \sum_{h=1}^L \sum_{h'=1}^L \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} W_h W_{h'} d_h d_{h'} \end{aligned} \quad (10.19)$$

性质 6 2^L 个统计量 $(y_{\alpha, a} - y_{\alpha})^2$ 的平均是 $V(y_{\alpha})$ 的一个无偏估计量。

$$E \left[\frac{1}{2^L} \sum_{\alpha=1}^{2^L} (y_{\alpha, a} - y_{\alpha})^2 \right] = V(y_{\alpha}) \quad (10.20)$$

三、平衡半样本估计

2^L 个统计量 $(y_{\alpha, a} - y_{\alpha})^2$ 的平均, 即 $\frac{1}{2^L} \sum_{\alpha=1}^{2^L} (y_{\alpha, a} - y_{\alpha})^2$ 是 $V(y_{\alpha})$ 的无偏估计。

然而, 当层数 L 较大时, 这个估计量的计算是不可行的。为了简化计算, 一个很自然的想法是选择一个小的半样本子集, 希望这个半样本子集尽量保留所有的信息, 这样既可达到简化计算的目的, 又能保证足够的精度。

假设这个半样本子集包含 k 个半样本, 由式(10.19)有

$$\begin{aligned} v_k(y_{\alpha}) &= \frac{1}{k} \sum_{\alpha=1}^k (y_{\alpha, a} - \bar{y}_{\alpha})^2 \\ &= v(\bar{y}_{\alpha}) + \frac{1}{2k} \sum_{h=1}^L \sum_{h'=1}^L \left[\sum_{\alpha=1}^k \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} \right] W_h W_{h'} d_h d_{h'} \end{aligned}$$

因此, 如果所选择的 k 个半样本对所有 $h < h' = 1, 2, \dots, L$ 都满足以下条件:

$$\sum_{\alpha=1}^k \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} = 0 \quad (10.21)$$

那么, $v_k(y_{\alpha})$ 就正好等于 $v(\bar{y}_{\alpha})$ 。因此, 我们说这 k 个半样本完全保留了 2^L 个半样本所包含的关于 $V(y_{\alpha})$ 的信息。满足条件式(10.21)的 k 组半样本称为平衡半样本。

如何确定平衡半样本呢? Plackett 和 Burman(1946) 给出了 $k \times k$ 阶正交矩阵 (k 为 4 的倍数, 为 Hadamard 矩阵) 的方法。例如, 表 10.5 和表 10.6 分别给出了 4

$\times 4$ 阶和 8×8 阶这样的矩阵, 其中列表示层, 行表示半样本; 在第 α 行第 h 列的位置上, $+1$ 表示层 h 的第一个单元被选入第 α 个半样本, -1 表示层 h 的第二个单元被选入第 α 个半样本。按这种方式定义的半样本即为平衡半样本。

在具体使用时, 可根据这种表的行确定半样本。例如若总体共分 5 层, 则使用表 10.6, 第一个半样本就是由第 2 行确定的: 在第 1, 2, 5 的位置上, 是 $+1$, 说明应在第 1, 2, 5 层中取第一个单元; 在第 3, 4 的位置上, 是 -1 , 表示在第 3, 4 层中应取第二个单元。至于哪个单元作为第一个单元, 哪个单元作为第二个单元, 可任意规定

表 10.5 和 10.6 还具有如下性质, 即除了最后一列外, 每列之和均为零。也就是说, 当 $L < k$ 时 ($k = 4$ 或 8), 有

$$\sum_{\alpha=1}^k \delta_h^{(\alpha)} = 0 \quad (10.22)$$

在 $L < k$ 时, 用 Plackett 和 Burman 方法选取的 k 组半样本都满足上式。因而在 $L < k$ 的条件下, 有

$$\sum_{\alpha=1}^k (y_{\alpha, \alpha} - y_{\alpha}) = \frac{1}{2} \sum_{h=1}^L W_h \left[\sum_{\alpha=1}^k \delta_h^{(\alpha)} \right] d_h = 0$$

$$\text{从而} \quad \frac{1}{k} \sum_{\alpha=1}^k y_{\alpha, \alpha} = y_{\alpha}, L < k \quad (10.23)$$

这与将所有的 2^L 个 $y_{\alpha, \alpha}$ 进行平均所得的结果完全一样。我们称同时满足式 (10.21) 和式 (10.22) 两个条件的半样本为完全正交平衡 (full orthogonal balance) 半样本。

如果 $L = k$, 由表 10.5 和表 10.6 可见, 最后一层均为 -1 , 不满足式 (10.22), 此时所抽选的半样本是平衡的但非完全正交平衡。要想抽取完全正交平衡半样本, k 的选择应该是大于 L 的 4 的最小整数倍。例如, 如果 $L = 8$, 则应取 $k = 12$ 。

表 10.5 2 层 — 4 层平衡半样本的确定

半样本	层			
	1	2	3	4
$\delta_h^{(1)}$	$+1$	$+1$	$+1$	-1
$\delta_h^{(2)}$	1	$+1$	1	1
$\delta_h^{(3)}$	1	1	$+1$	1
$\delta_h^{(4)}$	$+1$	1	-1	1

表 10.6

5 层—8 层平衡半样本的确定

半样本	层							
	1	2	3	4	5	6	7	8
$\delta_h^{(1)}$	+ 1	1	1	+ 1	- 1	+ 1	+ 1	- 1
$\delta_h^{(2)}$	+ 1	+ 1	+ 1	1	+ 1	1	+ 1	- 1
$\delta_h^{(3)}$	+ 1	+ 1	+ 1	1	- 1	+ 1	1	- 1
$\delta_h^{(4)}$	- 1	+ 1	+ 1	+ 1	1	1	+ 1	1
$\delta_h^{(5)}$	+ 1	1	1	+ 1	+ 1	1	1	1
$\delta_h^{(6)}$	- 1	+ 1	+ 1	+ 1	+ 1	+ 1	- 1	- 1
$\delta_h^{(7)}$	1	1	- 1	- 1	+ 1	+ 1	+ 1	- 1
$\delta_h^{(8)}$	1	1	1	1	- 1	- 1	- 1	1

四、部分平衡半样本

在复杂分层抽样方案中,层数 L 经常很大,即使平衡半样本方法已经减少了半样本数,但由于 $k \geq L$ 的要求,所需计算量仍然庞大。这时可以设计 k 组部分平衡半样本,具体方法如下。

假设有 L 层,采用 k 组半样本, $k < L$ 。假设 L 可以被 k 整除, $\frac{L}{k} = G$, F 是 L 层可分为 G 群。为叙述方便,假设 $L = 4$,按照平衡半样本方法,必须 $k \geq 4$ 。但这里我们取 $k = 2$,则 4 层分为 2 群,对包含第 1 层和第 2 层的第一群利用 2 阶 Hadamard 矩阵构造正交列,对包含第 3 层和第 4 层的第二群用同样方法,见表 10.7。

表 10.7

部分平衡半样本的确定

半样本	层			
	1	2	3	4
$\delta_h^{(1)}$	+ 1	+ 1	+ 1	+ 1
$\delta_h^{(2)}$	+ 1	- 1	+ 1	1

部分平衡半样本计算简单, k 组部分平衡半样本的方差估计量如下:

$$\sum_{\alpha=1}^k \frac{(y_{\alpha,\alpha} - \bar{y}_{\alpha})^2}{k} = \frac{1}{4} \sum_{h=1}^L W_h^2 (y_{h1} - y_{h2})^2 + \frac{1}{2} \sum_{h,j} W_h W_j (y_{h1} - y_{h2}) (y_{j1} - y_{j2}) \quad (10.24)$$

部分平衡半样本的方差估计量虽然不如完全平衡半样本精确,但也是无偏的。

五、用于多阶段抽样

以上介绍了适用层内放回简单随机抽样时的平衡半样本方法。这里将考虑使用不等概抽取的多阶段抽样的情形,假设在 L 层中的每一层初级抽样单元(PSU)都是按放回的 PPS 抽样抽取的。考虑总体总和 Y 的如下的无偏估计量:

$$\hat{Y} = \sum_{h=1}^L \hat{Y}_h = \sum_{h=1}^L \left(\frac{\hat{Y}_{h1}}{2z_{h1}} + \frac{\hat{Y}_{h2}}{2z_{h2}} \right) \quad (10.25)$$

式中, \hat{Y}_{hi} 为第 h 层第 i 个初级单元总和的一个无偏估计; z_{hi} 为第 h 层第 i 个初级单元每次抽取的概率。则 \hat{Y} 的通常的方差估计量为:

$$v(\hat{Y}) = \frac{1}{4} \sum_{h=1}^L \left(\frac{\hat{Y}_{h1}}{z_{h1}} + \frac{\hat{Y}_{h2}}{z_{h2}} \right)^2 \quad (10.26)$$

与放回的简单随机抽样情形一样,有 2^L 个可能的半样本,确定 k 个平衡半样本。对第 α 个半样本, Y 的估计量为:

$$\hat{Y}_\alpha = \sum_{h=1}^L \left(\delta_{h1\alpha} \frac{\hat{Y}_{h1}}{z_{h1}} + \delta_{h2\alpha} \frac{\hat{Y}_{h2}}{z_{h2}} \right) \quad (10.27)$$

其中, $\delta_{h1\alpha} = \begin{cases} 1, & \text{层 } h \text{ 中第 } 1 \text{ 个单元被选入第 } \alpha \text{ 个半样本} \\ 0, & \text{其他} \end{cases}$

$$\delta_{h2\alpha} = 1 - \delta_{h1\alpha}$$

\hat{Y} 的方差估计量为:

$$v_k(\hat{y}) = \frac{1}{k} \sum_{\alpha=1}^k (\hat{Y}_\alpha - \hat{Y})^2 \quad (10.28)$$

六、用于非线性估计

以上讨论中使用的都是线性估计量,下面介绍应用于非线性估计量的平衡半样本法。继续假设为放回 PPS 分层抽样设计,估计量 $\hat{\theta}$ 可能是比率、比率的差、回归系数、相关系数等。

令基于原始样本的比率估计量为 $\hat{\theta} = \frac{\hat{Y}}{\hat{X}}$, 其中的 \hat{Y} 和 \hat{X} 是 (10.25) 的形式,

则第 α 个半样本的估计为 $\hat{\theta}_\alpha = \frac{\hat{Y}_\alpha}{\hat{X}_\alpha} X$, 半样本估计量的均值为:

$$\hat{\theta} = \sum_{\alpha=1}^k \frac{\hat{\theta}_\alpha}{k}$$

对于非线性估计量, 一般 $\hat{\theta}$ 和 $\hat{\theta}$ 是不等的, 但多数调查实践中两者非常接近。

基于 k 个平衡半样本的 $V(\hat{\theta})$ 的估计量有以下几种形式可供选择:

1. 与线性问题类似的估计量:

$$v_k(\hat{\theta}) = \frac{1}{k} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 \quad (10.29)$$

2. 由于 k 个平衡半样本的余集也是平衡半样本, 因而也可以利用它们来构造方差估计量:

$$v_k^*(\hat{\theta}) = \frac{1}{k} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 \quad (10.30)$$

式中, $\hat{\theta}_\alpha$ 为基于第 α 个半样本的余集构造的估计量。

3. 结合式(10.29)和式(10.30)可得到另一个方差估计量:

$$v_k(\hat{\theta}) = \frac{1}{2} [v_k(\hat{\theta}) + v_k^*(\hat{\theta})] \quad (10.31)$$

4. 根据 k 组半样本与其余集的估计, $V(\hat{\theta})$ 的估计量:

$$v_k^*(\hat{\theta}) = \frac{1}{4k} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta}_\alpha)^2 \quad (10.32)$$

当 $\hat{\theta}$ 为线性估计时, $v_k(\hat{\theta}) = v_k^*(\hat{\theta}) = v_k(\hat{\theta}) = v_k^*(\hat{\theta})$;

若 $\hat{\theta}$ 为非线性估计, 它们一般不会相等。通常 $v_k(\hat{\theta})$, $v_k^*(\hat{\theta})$ 和 $v_k(\hat{\theta})$ 要比 $v_k^*(\hat{\theta})$ 大一些。

【例 10.3】拒答率调查

为研究被调查者拒答情况, 实施一项调查。抽样方式为分层随机抽样, 从三个城区中各自随机抽取两个居委会, 假设各层权重相同, 调查结果见表 10.8。试利用平衡半样本方法估计拒答率 \hat{R} 的方差。

表 10.8

样本的拒答情况

城 区	居委会 S_1		居委会 S_2	
	拒答户数(y_1)	合格调查户数(x_1)	拒答户数(y_2)	合格调查户数(x_2)
1	41	150	37	149
2	40	149	30	148
3	38	145	38	150
总 计	119	444	105	447

解:由于各层权重相同,拒答率的估计为:

$$\hat{R} = \frac{\sum_{h=1}^3 (y_{h1} + y_{h2})}{\sum_{h=1}^3 (x_{h1} + x_{h2})} = \frac{224}{891} = 0.251\ 402\ 92$$

用完全平衡半样本法估计拒答率 \hat{R} 的方差,抽样层数 $L = 3$,因为要求 $k \geq L$,因此取 $k = 4$,平衡半样本的确定见表 10.5。

取各区第一个居委会形成第一个半样本,该半样本及其余集的拒答率的估计为:

$$\hat{R}_1 = \frac{y_{11} + y_{21} + y_{31}}{x_{11} + x_{21} + x_{31}} = \frac{119}{444} = 0.268\ 018$$

$$\hat{R}_1' = \frac{y_{12} + y_{22} + y_{32}}{x_{12} + x_{22} + x_{32}} = \frac{105}{447} = 0.234\ 899$$

取第一区的第二个居委会、第二区的一个居委会以及第三区的第二个居委会形成第二个半样本,该半样本及其余集的拒答率的估计为:

$$\hat{R}_2 = \frac{y_{12} + y_{21} + y_{32}}{x_{12} + x_{21} + x_{32}} = \frac{115}{448} = 0.256\ 696$$

$$\hat{R}_2' = \frac{y_{11} + y_{22} + y_{31}}{x_{11} + x_{22} + x_{31}} = \frac{109}{443} = 0.246\ 049\ 7$$

取第一层和第二层的第二个居委会以及第三层第一个居委会形成第三个半样本,该半样本及其余集的拒答率的估计为:

$$\hat{R}_3 = \frac{y_{12} + y_{22} + y_{31}}{x_{12} + x_{22} + x_{31}} = \frac{105}{442} = 0.237\ 557$$

$$\hat{R}_3' = \frac{y_{11} + y_{21} + y_{32}}{x_{11} + x_{21} + x_{32}} = \frac{119}{449} = 0.265\ 033$$

取第一层第一个居委会、第二层和第三层的第二个居委会形成第四个半样本,

该半样本及其余集的拒答率的估计为:

$$\hat{R}_4 = \frac{y_{11} + y_{22} + y_{32}}{x_{11} + x_{22} + x_{32}} = \frac{109}{448} = 0.243\ 303\ 6$$

$$\hat{R}_4^c = \frac{y_{12} + y_{21} + y_{31}}{x_{12} + x_{21} + x_{31}} = \frac{115}{443} = 0.259\ 594$$

比是一个非线性统计量,下面用四种方法估计拒答率 \hat{R} 的方差:

$$v_k(\hat{R}) = \frac{1}{4} \sum_{a=1}^4 (\hat{R}_a - \hat{R})^2 = 0.000\ 141$$

$$v_k^c(\hat{R}) = \frac{1}{4} \sum_{a=1}^4 (\hat{R}_a^c - \hat{R})^2 = 0.000\ 138$$

$$v_k(\hat{R}) = \frac{1}{2} [v_k(\hat{R}) + v_k^c(\hat{R})] = 0.000\ 139$$

$$v_k^*(\hat{R}) = \frac{1}{4 \times 4} \sum_{a=1}^4 (\hat{R}_a - \hat{R}_a^c)^2 = 0.000\ 139$$

很明显,对这些数据来说, $v_k(\hat{R})$, $v_k^c(\hat{R})$ 和 $\bar{v}_k(\hat{R})$ 要比 $v_k^*(\hat{R})$ 之间的差别小。

【例 10.4】 铁路系统调查

铁路系统为估计收益—费用比,实施了一次货运调查。抽样框为运输记录文件,被分为 446 个层,从每层中抽出一个有 2 辆车的简单随机样本。根据样本数据,得总费用、总收益以及收益—费用比的估计值:

$$\hat{Y} = 11\ 758\ 070, \hat{X} = 18\ 266\ 375, \hat{R} = \frac{\hat{X}}{\hat{Y}} = 1.554$$

用平衡半样本法计算收益—费用比率估计量的方差。该调查中抽样层数 $L=446$,如采用完全平衡设计,需要 $k \geq L$;为方便计算并节省费用,最终采用部分平衡半样本法,取 $k=16$,部分平衡半样本的确定见表 10.9。

表 10.9 部分平衡半样本的确定

半样本	层													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$\delta_h^{(1)}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$\delta_h^{(2)}$	1	-1	1	-1	1	1	1	1	1	-1	1	-1	1	1
$\delta_h^{(3)}$	1	1	1	1	1	1	1	1	1	1	-1	-1	1	1
$\delta_h^{(4)}$	1	1	1	1	1	1	-1	1	1	-1	-1	1	1	1

续前表

半样本	层													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$\delta_h^{(5)}$	1	1	1	1	-1	1	1	1	1	1	1	1	-1	-1
$\delta_h^{(6)}$	1	1	1	1	1	1	-1	1	1	-1	1	-1	-1	1
$\delta_h^{(7)}$	1	1	1	-1	1	1	1	1	1	1	-1	-1	-1	-1
$\delta_h^{(8)}$	1	1	1	1	-1	1	1	1	1	1	1	1	1	1
$\delta_h^{(9)}$	1	1	1	1	1	1	1	1	1	1	-1	-1	-1	-1
$\delta_h^{(10)}$	1	-1	-1	-1	1	-1	1	-1	-1	1	-1	1	1	1
$\delta_h^{(11)}$	1	1	1	1	1	1	-1	-1	-1	1	1	1	1	1
$\delta_h^{(12)}$	1	1	1	1	1	1	1	1	-1	1	1	1	1	1
$\delta_h^{(13)}$	1	1	1	1	1	1	1	-1	-1	-1	1	1	1	1
$\delta_h^{(14)}$	1	-1	-1	-1	-1	1	-1	1	1	1	-1	1	1	1
$\delta_h^{(15)}$	1	1	1	1	1	-1	1	1	-1	-1	1	1	1	1
$\delta_h^{(16)}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1

将 446 个层分为 14 组, 每组含有 32 个抽样层(最后一组含 30 个抽样层)。每一组内各抽样层采用相同的半样本选取方案, 比如按照第 16 种半样本方案, 第一组的 32 层都选取第一个单元, 第二组的 32 层都选取第二个单元, 第三组的 32 层都选取第一个单元, 依次类推, 形成第 16 个半样本。最终得到 16 个半样本的估计值如表 10.10。

表 10.10 半样本估计值

子样本	收益 - 费用比(\hat{R}_s)	$(\hat{R}_s - \hat{R})^2$
1	1.54	0.000 196
2	1.53	0.000 576
3	1.55	0.000 016
4	1.53	0.000 576
5	1.55	0.000 016
6	1.56	0.000 036
7	1.53	0.000 576

$$\hat{\theta}_\alpha = k\hat{\theta} - (k-1)\hat{\theta}_{(\alpha)} \quad (10.33)$$

$\hat{\theta}$ 通常包含全样本各随机观测值提供的关于 θ 的全部信息, 相应的 $\hat{\theta}_{(\alpha)}$ 包含除去第 α 组子样本外其余所有随机观测值提供的关于 θ 的全部信息。从上式可以看出这样的含义: $\hat{\theta}_\alpha$ 可以看做从 $\hat{\theta}$ 所包含的 θ 的信息中剔除了 $\hat{\theta}_{(\alpha)}$ 关于 θ 的信息, 因而虚拟值 $\hat{\theta}_\alpha$ 可以看做仅仅包含第 α 组子样本所提供的关于 θ 的信息。因此, 虚拟值 $\hat{\theta}_\alpha$ 可以近似看成独立同分布

θ 的刀切法估计定义为所有的 $\hat{\theta}_\alpha$ 的平均值

$$\hat{\theta} = \frac{1}{k} \sum_{\alpha=1}^k \hat{\theta}_\alpha \quad (10.34)$$

而 $\hat{\theta}$ 的刀切法方差估计为:

$$v_1(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 \quad (10.35)$$

实践中, $v_1(\hat{\theta})$ 不仅用于估计 $\hat{\theta}$ 的方差, 也用于估计 $\hat{\theta}$ 的方差 $V(\hat{\theta})$, 即

$$v_1(\hat{\theta}) = v_1(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 \quad (10.36)$$

另外, 对 $\hat{\theta}$ 的方差 $V(\hat{\theta})$ 的估计还可以使用:

$$v_2(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 \quad (10.37)$$

相对于 $v_1(\hat{\theta})$ 而言, $v_2(\hat{\theta})$ 是一个保守的估计。

二、有限总体的刀切法估计

应用刀切法进行有限总体的方差估计之前, 应该先将原始样本划分为 k 个随机组, 这些随机组可以是独立随机组, 也可以是非独立随机组。

(一) 放回的简单随机抽样

假设总体单元 Y_1, Y_2, \dots, Y_N , 待估参数为总体均值 Y , 从总体中抽取一个样本量为 n 的有放回的简单随机样本 y_1, y_2, \dots, y_n , $y = \sum_{i=1}^n \frac{y_i}{n}$ 是 Y 的无偏估计。其方差 $V(y) = \sum_{i=1}^n \frac{(Y_i - Y)^2}{nN}$ 具有无偏估计:

$$v(y) = \sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n(n-1)}$$

应用刀切法,将样本分成大小为 m 的 k 个随机组, $n = mk$ 。由于抽样是放回的,随机组之间是独立的。取 $\hat{\theta} = \bar{y}$ 。由于其为线性形式,故总体均值 Y 的刀切法估计量为:

$$\hat{\theta} = \frac{1}{k} \sum_{a=1}^k \hat{\theta}_a = ky - (k-1) \sum_{a=1}^k \frac{y_{(a)}}{k} \quad (10.38)$$

式中, $y_{(a)} = \frac{1}{m(k-1)} \sum_{i=1}^{m(k-1)} y_i$ 为舍弃第 a 组观测值后的样本均值。

$$v_1(\hat{\theta}) = k \frac{1}{k} \sum_{a=1}^k (y_{(a)} - y)^2 \quad (10.39)$$

很容易验证

$$\hat{\theta} = y$$

$$E[v_1(\hat{\theta})] = V(\hat{\theta}) = V(y)$$

当且仅当 $k = n, m = 1$ 时, $v(\hat{\theta}) = v(y)$ 。

(二) 放回的 PPS 抽样

假设按放回的 PPS 抽样方式抽取一个样本量为 n 的样本,第 j 个单元每次入样的概率为 z_j ,则总体总和 Y 的估计及其方差为:

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i}$$

$$V(\hat{Y}) = \sum_{i=1}^n \frac{z_i \left(\frac{y_i}{z_i} - Y \right)^2}{n}$$

$V(\hat{Y})$ 的无偏估计为:

$$v(\hat{Y}) = \sum_{i=1}^n \frac{\left(\frac{y_i}{z_i} - \hat{Y} \right)^2}{n(n-1)}$$

应用刀切法,取 $\hat{\theta} = \hat{Y}$,假定 $n = mk$,则第 a 个虚拟值为:

$$\hat{\theta}_a = k \hat{\theta} - (k-1) \hat{\theta}_{(a)}$$

式中, $\hat{\theta}_{(a)} = \frac{1}{m(k-1)} \sum_{i=1}^{m(k-1)} \frac{y_i}{z_i} = \hat{Y}_{(a)}$ 是舍弃第 a 组观测值后的估计量。于是, Y

的刀切法估计为:

$$\hat{\theta} = \frac{1}{k} \sum_{\alpha=1}^k \hat{\theta}_{\alpha} = \frac{1}{k} \sum_{\alpha=1}^k \hat{Y}_{(\alpha)} \quad (10.40)$$

$\hat{\theta}$ 的方差 $V(\hat{\theta})$ 的刀切法估计则为:

$$v_1(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_{\alpha} - \hat{\theta})^2 \quad (10.41)$$

不难验证:

$$\hat{\theta} = \hat{Y}$$

$$E[v_1(\hat{\theta})] = V(\hat{\theta}) = V(\hat{Y})$$

容易看出 $v_1(\hat{\theta})$ 一般并不等于通常的方差估计量。但当 $k = n, m = 1$ 时, 有

$$v_1(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{\alpha=1}^n (y_{\alpha} - \bar{y})^2$$

这与通常的方差估计量一致。

(三) 不放回的简单随机抽样

假设采用不放回简单随机抽样抽取一个样本量为 n 的样本。将该样本分成大小为 m 的 k 个随机组, $n = mk$ 。由于抽样是不放回的, 随机组之间就不独立了。取 $\hat{\theta} = \bar{y}$ 。由于其为线性形式, 故总体均值 Y 的刀切法估计量即为其本身:

$$\hat{\theta} = \frac{1}{k} \sum_{\alpha=1}^k \hat{\theta}_{\alpha} = \bar{y} \quad (10.42)$$

式中, 第 α 个虚拟值 $\hat{\theta}_{\alpha}$ 定义为:

$$\hat{\theta}_{\alpha} = k\hat{\theta} - (k-1)\hat{\theta}_{(\alpha)} = ky - (k-1)y_{(\alpha)}$$

而 $y_{(\alpha)} = \frac{1}{m(k-1)} \sum_{i=1}^{m(k-1)} y_i$ 表示舍弃第 α 组观测值后的样本均值。

$\hat{\theta}$ 的刀切法方差估计为:

$$v_1(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_{\alpha} - \hat{\theta})^2 = \frac{1}{k} \sum_{\alpha=1}^k (y_{(\alpha)} - y)^2 \quad (10.43)$$

$v_1(\hat{\theta})$ 对于放回的简单随机抽样是 $V(\hat{\theta})$ 的无偏估计, 但在不放回简单随机抽样下, $v_1(\hat{\theta})$ 不再是 $V(\hat{\theta})$ 的无偏估计了。事实上, 可以证明

$$E[v_1(\hat{\theta})] = \frac{S^2}{n} \quad (10.44)$$

式中, S^2 为总体方差。因此, $v_1(\hat{\theta})$ 的偏倚为:

$$E[v_1(\hat{\theta})] - V(\hat{\theta}) = \frac{fS^2}{n} \quad (10.45)$$

若抽样比 f 可忽略, 则 $v_1(\hat{\theta})$ 是近似无偏的; 如果抽样比 f 不能忽略, 则一个自然的修正是采用估计量 $(1-f)v_1(\hat{\theta})$ 作为方差 $V(\hat{\theta})$ 的估计。另一个修正的办法是将虚拟值定义为:

$$\hat{\theta}_a^* = k\hat{\theta} - (k-1)\hat{\theta}_{(a)} \quad (10.46)$$

式中, $\hat{\theta}_{(a)} = y + (1-f)^{\frac{1}{2}}(y_{(a)} - y)$ 。此时 Y 的刀切法估计为:

$$\hat{\theta}^* = \frac{1}{k} \sum_{a=1}^k \hat{\theta}_a^* \quad (10.47)$$

$V(\hat{\theta}^*)$ 的刀切法估计则为:

$$v_1(\hat{\theta}^*) = \frac{1}{k(k-1)} \sum_{a=1}^k (\hat{\theta}_a^* - \hat{\theta}^*)^2 \quad (10.48)$$

对于修正的刀切法估计:

$$\hat{\theta}^* = y$$

$$E[v_1(\hat{\theta}^*)] = V(y) = \frac{1-f}{n} S^2$$

当且仅当 $k = n, m = 1$ 时, $v_1(\hat{\theta}^*) = v(y)$ 。

(四) 用于比率估计

刀切法的用途主要是对复杂样本或非线性估计进行方差估计。这里, 给出它在比值估计中的应用, 不限定具体是什么抽样方案。

假定要估计比值 $R = \frac{Y}{X}$, 其中 Y 与 X 是总体总和。通常的估计量是 $\hat{R} = \frac{\hat{Y}}{\hat{X}}$,

而 \hat{Y}, \hat{X} 是基于特定抽样方案的总体总和的估计。将样本分成大小为 m 的 k 个随机组, $n = mk$, 则虚拟值定义为:

$$\hat{R}_a = k\hat{R} - (k-1)\hat{R}_{(a)} \quad (10.49)$$

式中, $\hat{R}_\alpha = \frac{\hat{Y}_{(\alpha)}}{\hat{X}_{(\alpha)}}$, 而 $\hat{Y}_{(\alpha)}, \hat{X}_{(\alpha)}$ 分别是舍弃第 α 个随机组后 Y 与 X 的估计。由此

得出 R 的刀切法估计为:

$$\hat{R} = \frac{1}{k} \sum_{\alpha=1}^k \hat{R}_\alpha \quad (10.50)$$

而 \hat{R} 或 \hat{R}_α 的刀切法方差估计为:

$$v_1 = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{R}_\alpha - \hat{R})^2 \quad (10.51)$$

$$v_2 = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{R}_\alpha - \hat{R})^2 \quad (10.52)$$

【例 10.5】 继续使用例 10.3 的拒答率调查数据, 利用刀切法估计拒答率 \hat{R} 的方差。

解: 根据样本数据, 估计拒答率 \hat{R} :

$$\hat{R} = \frac{\sum_{i=1}^3 \sum_{j=1}^2 y_{ij}}{\sum_{i=1}^3 \sum_{j=1}^2 x_{ij}} = \frac{224}{891} = 0.251\ 402\ 92$$

根据刀切法估计拒答率 \hat{R} 的方差。根据抽样层, 样本分为 3 组, 有

$$\hat{R}_{(1)} = \frac{\sum_{j=1}^2 y_{2j} + \sum_{j=1}^2 y_{3j}}{\sum_{j=1}^2 x_{2j} + \sum_{j=1}^2 x_{3j}} = \frac{146}{592} = 0.246\ 621\ 6$$

$$\hat{R}_{(2)} = \frac{\sum_{j=1}^2 y_{1j} + \sum_{j=1}^2 y_{3j}}{\sum_{j=1}^2 x_{1j} + \sum_{j=1}^2 x_{3j}} = \frac{154}{594} = 0.259\ 259$$

$$\hat{R}_{(3)} = \frac{\sum_{j=1}^2 y_{2j} + \sum_{j=1}^2 y_{1j}}{\sum_{j=1}^2 x_{2j} + \sum_{j=1}^2 x_{1j}} = \frac{148}{596} = 0.248\ 322\ 1$$

因此虚拟值为:

$$\hat{R}_1 = 3\hat{R} - (3-1)\hat{R}_{(1)} = 0.260\ 966$$

$$\hat{R}_2 = 3\hat{R} - (3-1)\hat{R}_{(2)} = 0.23569$$

$$\hat{R}_3 = 3\hat{R} - (3-1)\hat{R}_{(3)} = 0.257564$$

根据刀切法,估计拒答率:

$$\hat{R} = \frac{1}{3}(\hat{R}_1 + \hat{R}_2 + \hat{R}_3) = 0.251407$$

用刀切法估计拒答率 \hat{R} 的方差:

$$\begin{aligned} v_1(\hat{R}) &= \frac{1}{k(k-1)} \sum_{a=1}^k (\hat{R}_a - \hat{R})^2 \\ &= \frac{1}{3(3-1)} \sum_{a=1}^3 (\hat{R}_a - 0.251407)^2 = 0.00006272 \\ v_2(\hat{R}) &= \frac{1}{k(k-1)} \sum_{a=1}^k (\hat{R}_a - \hat{R})^2 \\ &= \frac{1}{3(3-1)} \sum_{a=1}^3 (\hat{R}_a - 0.25140292)^2 = 0.000062.72 \end{aligned}$$

§ 10.5 泰勒级数法

以上几节介绍的随机组估计法、平衡半样本法和刀切法都采用样本复制的原理对复杂样本进行方差估计,本节介绍的泰勒级数法(Taylor series method)是一种线性化方法,主要是利用泰勒展开的办法用线性估计去逼近非线性估计,由此给出非线性估计量方差的一个近似估计。显然,泰勒级数法主要是针对非线性估计量的方差估计,而且它本身不能独自地用于方差估计,在提供了非线性估计量的线性近似之后,还需要结合其他方法去估计这个线性近似的方差。

对于一个有限总体 N ,令 $Y = (Y_1, Y_2, \dots, Y_p)$ 表示总体参数的一个 p 维向量,以 $\hat{Y} = (\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_p)$ 表示基于样本量为 n 的样本的相应估计量。估计量 \hat{Y}_i 的形式取决于抽样设计。在大多数应用中, Y_i 表示 p 个不同的调查指标的总体总和或总体均值,这时, \hat{Y}_i 通常是 Y_i 的无偏估计。

假定要估计的总体参数不是 Y ,而是 Y 的函数形式 $\theta = g(Y)$,相应的估计量应为 $\hat{\theta} = g(\hat{Y})$ 。我们面临两个问题:(1)找到 $\hat{\theta}$ 的方差的近似表达式;(2)构造 $\hat{\theta}$ 方差的合适的估计量。

如果函数 $g(y)$ 在包含 Y 和 \hat{Y} 的某个开集内具有连续的二阶偏导数, 则将 $\hat{\theta}$ 在 Y 处泰勒展开, 得

$$\hat{\theta} - \theta = \sum_{i=1}^p \frac{\partial g(Y)}{\partial y_i} (\hat{Y}_i - Y_i) + \frac{1}{2} \sum_{i=1}^p \sum_{j=1}^p \frac{\partial^2 g(\tilde{Y})}{\partial y_i \partial y_j} (\hat{Y}_i - Y_i)(\hat{Y}_j - Y_j) \quad (10.53)$$

式中, \tilde{Y} 位于 \hat{Y} 与 Y 之间。

在有限总体中, 一般认为(10.53)式中的第二项相对于第一项来说是可忽略的, 因而近似地有

$$\hat{\theta} - \theta \approx \sum_{j=1}^p \frac{\partial g(Y)}{\partial y_j} (\hat{Y}_j - Y_j) \quad (10.54)$$

$\hat{\theta}$ 的均方误差近似为:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &\approx E \left[\sum_{j=1}^p \frac{\partial g(Y)}{\partial y_j} (\hat{Y}_j - Y_j) \right]^2 \\ &= \sum_{i=1}^p \sum_{j=1}^p \frac{\partial g(Y)}{\partial y_i} \cdot \frac{\partial g(Y)}{\partial y_j} \text{cov}(\hat{Y}_i, \hat{Y}_j) \\ &= \hat{d}' \sum d' \end{aligned}$$

式中, $\sum = V(\hat{Y})$ 为 \hat{Y} 的协方差矩阵; d 为 p 维向量, 其第 j 个元素为 $d_j = \frac{\partial g(Y)}{\partial y_j}$ 。

至于上述均方误差的估计, 只需将相应的样本估计代入即可。这样, $\text{MSE}(\hat{\theta})$ 的估计量为:

$$\text{MSE}(\hat{\theta}) = \hat{d}' \sum \hat{d}$$

式中, \sum 为 \sum 的估计; \hat{d} 的元素为 $\hat{d}_j = \frac{\partial g(\hat{Y})}{\partial y_j}$ 。

对于一阶近似来说, 方差 $V(\hat{\theta})$ 与偏倚 $B(\hat{\theta})$ 一般具有相同的阶。故 $[B(\hat{\theta})]^2$ 相对于 $V(\hat{\theta})$ 来说具有更高的阶, 因而可忽略, 即有

$$\text{MSE}(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2 \approx V(\hat{\theta})$$

当然, 如果进一步泰勒展开, 可以得到二阶或更高阶的近似, 但对于一般的复

杂样本调查的方差估计,一阶近似已可以产生较满意的结果。此外,若总体发生严重偏倚,这种近似是不可靠的

§ 10.6 方法的比较

本章介绍了复杂样本方差估计的四种方法,即随机组方法、平衡半样本方法、刀切法以及泰勒级数法。在实际应用中,究竟应该选用哪一种方法?本节从方差估计的精度、费用和时间以及操作的简便性等方面对这些方法进行比较。

(一) 精度

首先,应注意方差估计量的精度可以用不同的标准进行衡量,如偏倚、均方误差、置信区间覆盖概率等。不同的方差估计量在不同的标准下可能都是最好的。根据 K. M. Wolter 的研究,用偏倚和均方误差标准都难以直接判断方差的最优估计量,一般把置信区间覆盖概率作为最重要的精度标准。

从精度上考虑,四种方法在大样本情况下效果差不多。对于中小样本,已有的蒙特卡洛研究的结果显示,若以偏倚和均方误差作为标准,则泰勒级数法较好,在某些情况下可能是最好的;而随机组方差估计在许多应用中较之其他三种方法有更大的均方误差,但是从置信区间的覆盖概率的角度看,平衡半样本方法最好,其次是随机组方法和刀切法。

(二) 费用和时间

从费用、时间上考虑,随机组方法和平衡半样本方法都是值得推荐的。这两种方法的计算都有现成的软件,数据处理费用相对较低。在大规模的调查中,平衡半样本方法的费用还要低。刀切法比较费时费力,主要是因为目前还没有现成的软件来应用这一方法。如前所述,泰勒级数法本身不能单独使用,它必须与其他方法结合起来才能对方差进行估计,其费用的高低主要依赖于与之配合使用的其他方差估计方法。例如,若用刀切法估计协方差矩阵,那么泰勒级数法的费用可能相当大。

(三) 操作的简便性

一般地说,随机组方法是最灵活的方差估计方法之一,适用于几乎任何估计量;同时,它也是用途最广的方法,适用于几乎任何抽样设计。平衡半样本法从适用的估计量的类型看,其灵活性不逊于随机组法,但是从抽样设计的角度看,它常常被认为局限于分层的、每层抽两个单元的抽样设计。当然,使用更复杂的平衡方案,平衡半样本法也可用于每层抽三个及以上单元,或者每层只抽一个单元的抽样设计。刀切法可用于抽样调查实践中可能出现的大多数估计量,从应用的广度看,它

与平衡半样本法不相上下,但比不上随机组方法。泰勒级数法在适用抽样设计和估计量上与其他方法有同样的灵活性。

总之,对方差估计方法进行选择是一个复杂的问题,需要综合考虑精度、费用和时间、可操作性等各种因素,在它们之间进行权衡。

小 结

实际调查往往是一种复杂抽样调查,对于复杂样本的方差估计需要采用随机组方法、平衡半样本法、刀切法以及泰勒级数法等方差估计方法。

随机组方法的实质是按一定的抽样方案从总体中抽取若干组样本,对于每一组样本都建立有关参数的估计量,这些估计量之间的离散程度,即样本方差就可用于计算全样本估计量的方差。

平衡半样本方法也是一种重抽样方法,它将各层中随机组数减为两个以提高方差估计计算的效率。但是它与随机组方法有所区别。

刀切法用重抽样技巧可以将原来的总体进行复制,在复制的总体中,可以使用原来的抽样办法再复制抽样样本及构造同样结构的有关参数的统计量。由于复制的总体及统计量是原有总体及统计量的一个缩影,而在复制的模型中,包括统计量的均值、方差等特性在内的几乎一切为我们所关心的指标均可以通过计算得到。

泰勒级数法属于线性化的方差估计方法。其实质是将非线性估计线性化。利用泰勒级数展开可以用线性估计去逼近非线性估计量,从而得到非线性方差估计量的近似估计。

对复杂样本的方差估计方法进行选择是一个复杂的问题,需要综合考虑精度、费用和时间、可操作性等各种因素,在它们之间进行权衡。

本章附录 复杂样本的方差估计的性质证明

1. 证明性质 1: 设 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ 是相互独立的随机变量,并且具有相同的期望 $E(\hat{\theta}_a) = \mu$, 定义 $\hat{\theta} = \frac{1}{k} \sum_{a=1}^k \hat{\theta}_a$, 则 $E(\hat{\theta}) = \mu$, $\hat{\theta}$ 的方差 $V(\hat{\theta})$ 的一个无偏估计是

$$v(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{a=0}^k (\hat{\theta}_a - \hat{\theta})^2 \quad (10.2)$$

证明:显然, $E(\hat{\theta}) = \mu$, 而 $v(\hat{\theta})$ 可以表达为:

$$v(\hat{\theta}) = \frac{1}{k(k-1)} \left(\sum_{a=1}^k \hat{\theta}_a^2 - k\hat{\theta}^2 \right)$$

$\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ 是相互独立的, 故有

$$\begin{aligned} E[v(\hat{\theta})] &= \frac{1}{k(k-1)} \left[\sum_{a=1}^k [V(\hat{\theta}_a) + \mu^2] - k[V(\hat{\theta}) + \mu^2] \right] \\ &= \frac{1}{k(k-1)} [k^2 V(\hat{\theta}) - kV(\hat{\theta})] \\ &= V(\hat{\theta}) \end{aligned}$$

统计量 $\hat{\theta}$ 可作为 θ 的估计量, 而 $v(\hat{\theta})$ 是方差 $V(\hat{\theta})$ 的随机组估计量。

2. 证明性质 3: 设 $E(\hat{\theta}_a) = \mu_a$ (μ_a 不一定等于 θ), 则

$$E(\hat{\theta}) = \frac{1}{k} \sum_{a=1}^k \mu_a = \mu$$

$$\begin{aligned} \text{且 } E[v(\hat{\theta})] &= V(\hat{\theta}) + \frac{1}{k(k-1)} \sum_{a=1}^k (\mu_a - \mu)^2 - \frac{2}{k(k-1)} \sum_{a=1}^k \sum_{\beta > a}^k \text{Cov}(\hat{\theta}_a, \hat{\theta}_\beta) \\ &\quad (10.11) \end{aligned}$$

证明: 显然, $E(\hat{\theta}) = \mu$, $v(\hat{\theta})$ 可以表示成

$$\begin{aligned} v(\hat{\theta}) &= \hat{\theta}^2 - \frac{2}{k(k-1)} \sum_{a=1}^k \sum_{\beta > a}^k \hat{\theta}_a \hat{\theta}_\beta \\ E[v(\hat{\theta})] &= E(\hat{\theta}^2) - \frac{2}{k(k-1)} \sum_{a=1}^k \sum_{\beta > a}^k E[\hat{\theta}_a \hat{\theta}_\beta] \\ &= [V(\hat{\theta}) + \mu^2] - \frac{2}{k(k-1)} \sum_{a=1}^k \sum_{\beta > a}^k [\text{Cov}(\hat{\theta}_a, \hat{\theta}_\beta) + \mu_a \mu_\beta] \\ &= V(\hat{\theta}) + \frac{1}{k(k-1)} \sum_{a=1}^k (\mu_a - \mu)^2 \\ &\quad - \frac{2}{k(k-1)} \sum_{a=1}^k \sum_{\beta > a}^k \text{Cov}(\hat{\theta}_a, \hat{\theta}_\beta) \end{aligned}$$

3. 证明性质 6: 2^L 个统计量 $(y_{st,a} - y_{..})^2$ 的平均是 $V(y_a)$ 的一个无偏估计量:

$$E\left[\frac{1}{2^L} \sum_{a=1}^L (y_{st,a} - y_{..})^2\right] = V(y_a) \quad (10.20)$$

证明:定义

$\delta_h^{(\alpha)} = 2\delta_{h1\alpha} - 1 \begin{cases} = 1, \text{第 } h \text{ 层中第一个单元被选入第 } \alpha \text{ 个半样本} \\ = -1, \text{第 } h \text{ 层中第二个单元被选入第 } \alpha \text{ 个半样本} \end{cases}$

则

$$y_{st,a} - y_{st} = \frac{1}{2} \sum_{h=1}^L W_h \delta_h^{(\alpha)} d_h$$

由于 $\delta_h^{(\alpha)2} = 1$, 于是

$$\begin{aligned} (y_{st,a} - y_{st})^2 &= \frac{1}{4} \sum_{h=1}^L W_h^2 d_h^2 + \frac{1}{2} \sum_{h=1}^L \sum_{h' > h}^L \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} W_h W_{h'} d_h d_{h'} \\ &= v(y_{st}) + \frac{1}{2} \sum_{h=1}^L \sum_{h' > h}^L \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} W_h W_{h'} d_h d_{h'} \end{aligned}$$

因为层 h 和层 h' 中任何一对单元都是正好出现在 2^{L-2} 个半样本中, 故有

$\sum_{\alpha=1}^{2^L} \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} = 0$, 则

$$\begin{aligned} &\sum_{\alpha=1}^{2^L} (y_{st,a} - y_{st})^2 \\ &= 2^L v(y_{st}) + \frac{1}{2} \sum_{h=1}^L \sum_{h' > h}^L \left[\sum_{\alpha=1}^{2^L} \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} \right] W_h W_{h'} d_h d_{h'} \\ &= 2^L v(y_{st}) \end{aligned}$$

因此

$$E \left[\frac{1}{2^L} \sum_{\alpha=1}^{2^L} (\bar{y}_{st,a} - y_{st})^2 \right] = E[v(y_{st})] = V(\bar{y}_{st})$$

习 题

1. 为了估计某镇的失业率, 进行分层随机抽样, 四个城区作为四层, 并假定层权 $W_h = \frac{1}{4} (h = 1, 2, 3, 4)$; 独立抽取 3 个样本, 分别调查劳动人口数及失业人数, 结果见下表 (x_{hi} 表示劳动人口数, y_{hi} 表示失业人数), 试估计该镇失业率并估计其方差。

某镇失业情况调查结果

层	样本 1		样本 2		样本 3	
	x_{h1}	y_{h1}	x_{h2}	y_{h2}	x_{h3}	y_{h3}
1	520	46	515	50	515	46
2	501	37	488	46	488	40
3	579	58	469	43	469	47
4	507	40	500	49	496	44

2. 假设某学校调查学生对学校伙食的满意度, 抽样方式为分层随机抽样, 从三个年级中各自随机抽取两个班级, 假设各年级权重相同, $W_h = \frac{1}{3} (h = 1, 2, 3)$; 调查结果表示满意的学生人数见下表, 试估计学生对学校伙食的满意比率 \hat{R} 并用平衡半样本方法估计其方差。

学生对学校伙食的满意情况

年 级	班级(S_1)		班级(S_2)	
	满意人数(y)	就餐人数(r_1)	满意人数(y_2)	就餐人数(r_2)
1	41	80	37	79
2	40	79	30	68
3	38	72	38	55
总 计	119	231	99	202

3. 假设某镇有 10 个街道, 每个街道有 15 个居委会。为了调查该镇的人口出生率, 采用二阶简单随机抽样法抽取 4 个街道, 并在每个街道中抽取 6 个居委会。对每个被抽中的居委会调查其上一年的户口数及新生婴儿数, 结果见下表(x 表示人口数, y 表示新生婴儿数)。试估计该镇的人口出生率并给出方差估计。

某镇人口数及新生婴儿数抽样结果

街道 \ 居委会	1		2		3		4		5		6	
	x	y	x	y	x	y	x	y	x	y	x	y
1	520	1	485	2	496	10	515	3	518	9	492	7
2	405	5	501	6	517	3	520	5	482	2	488	4
3	504	4	532	7	579	6	462	7	469	3	519	4
4	529	5	530	2	425	1	523	10	567	7	527	3



第 11 章

调查中的非抽样误差

抽样调查中的误差包括抽样误差和非抽样误差。抽样误差是指由于抽样的随机性所引起的样本统计量的数值与总体目标变量真值之间的差异,它是抽样调查所特有的。抽样误差在概率抽样的条件下可以计量,并通过抽样设计加以控制。前面各章的内容主要是围绕抽样误差的计量和控制展开的,当然这是假定样本的数据是可以准确获得的。事实上抽样调查中除了抽样误差以外,还存在大量的非抽样误差,本章就讨论这个问题。第一节是对非抽样误差的综合性介绍,第二节讨论抽样框误差,第三节讨论无回答误差,第四节讨论计量误差,第五节讨论离群值的检测和处理。

§ 11.1 引言

非抽样误差是指除抽样误差以外的,由于各种原因引起的误差。在概率抽样、非概率抽样、其他全面调查和非全面调查以及普查中,非抽样误差都有可能存在。

同抽样误差相比,非抽样误差有如下特点:

首先,非抽样误差不是由于抽样的随机性带来的,所以在抽样调查中,它不可能随着样本量的增大而变小。有时情况可能还会相反,样本量越大,非抽样误差也越大,因为随着调查中更多人员的涉入,会增大非抽样误差出现的机会。

其次,在抽样调查中,由于非抽样误差的影响,往往造成估计量的有偏。例如,如果非抽样误差产生于调查中的无回答,而回答层和无回答层的被调查单元在目标变量方面存在差异,仅仅用回答层的观测数据对总体目标变量进行推断,就会造成有偏估计。

第二,有些非抽样误差难以识别和测定。例如,如果抽样框是不完善的,而调查设计人员并没有意识到,由不完善的抽样框进行设计所得到的调查结果自然包含非抽样误差,而使用数据人员却没有意识到,也不可能知道。另一种情况是,调查人员意识到非抽样误差可能存在,但无法准确判断,无法对其进行计量。在调查实践中后一种情况更为多见。

最后,由于产生非抽样误差的渠道众多,有些非抽样误差成因复杂,尤其当调查对象是人的时候,社会因素、经济因素对非抽样误差的范围和程度都产生不可忽视的影响。而且与抽样误差相比,对非抽样误差的研究尚有距离。因此,从实践角度看,非抽样误差对调查数据质量和估计结果的负面影响是非常大的,对此必须引起高度重视。

非抽样误差可以产生于抽样调查的各个阶段,包括调查及抽样设计、数据采集及数据的处理与分析阶段。

1. 调查及抽样设计阶段。调查设计包括多项工作,哪一项工作出现问题都可能造成难以补救的后果。例如,调查的问卷设计有缺陷,所用词汇的含义不清,造成被调查者的多种理解而提供了不准确的信息。抽样设计中,抽样框不完善是一个主要问题。不完善的抽样框是指抽样框中包含的单元与目标总体中的单元不一致,例如属于调查对象的单元在抽样框中不存在,不属于调查对象的单元却出现在抽样框中。不完善抽样框还包括这样的情况,即抽样框中的辅助信息与现实情况严重偏离,造成样本抽选的“误导”,使用不完善的抽样框是产生非抽样误差的一个重要原因。

2. 数据采集阶段。这又可以分为两个方面,一个方面是调查实施过程中,调查人员没有从被调查者那里得到所需要的信息,这种情况的产生可能有多种原因。例如由于地址不详或搬迁,调查人员没有找到被调查者,或者被调查者不在家,或者被调查者虽然在家,却由于某种原因没有接受调查。这种现象通常被称为无回答。无回答是造成数据采集阶段非抽样误差的主要原因。另一个方面是在数据采集过程中,被调查者虽然提供了回答,但与真实情况不一致。这种情况大多在敏感性调

查项目上出现。如果调查实施后发现被调查者提供的信息明显失真而将其剔除,这就变成了无回答

实践中调查数据的失真主要来自于被调查者,但有时也与调查人员有关,如调查人员有意或无意的诱导,记录调查结果出现错误等。当被调查单元是物体时,计量工具不精确也会使测量结果出现误差

3. 数据处理与分析阶段。主要指对调查资料进行审核、整理、编码和录入过程中出现差错所引起的误差。误差还可能产生于不正确的估计程序之中,例如应当加权却没有加权,或者使用与抽样方式不相匹配的估计方式,等等。对于最后一种类型的误差,将其归入调查设计误差也是可以的。

上述非抽样误差按其来源、性质不同,可以分为以下三类:

- (1) 抽样框误差,即由不完善的抽样框引起的误差。
- (2) 无回答误差,即由于种种原因没有从被调查单元获得调查结果,造成调查数据的缺失。
- (3) 计量误差,即所获得的调查数据与其真值之间不一致造成的误差。

§ 11.2 抽样框误差

一、概念

为了说明抽样框误差,有必要对总体的概念重新说明。抽样调查中的总体有两个,一个是目标总体,即作为调查研究对象的全体,这是通常意义上所说的总体;另一个是抽样总体,即从中抽选样本的总体。抽样总体的具体表现是抽样框。理想抽样框的标志是目标总体和抽样总体完全重合,就是说目标总体单元和抽样总体单元完全是一一对应的关系。否则,抽样框就是不完善的,这意味着有可能出现抽样框误差。

抽样调查中有一个完善的抽样框当然最好,但在实践中由于种种原因,特别是由于资料方面的原因,构造出完善的抽样框往往不容易。不完善抽样框的主要问题是总体中单元数 N 不准确,这时利用样本统计量对总体参数进行估计就可能产生估计偏倚。这种误差并不是来自于抽样的随机性,而是产生于不完善的抽样框,所以抽样框误差是一种非抽样误差。

对抽样框误差进行分析,首先是把握抽样框误差的类型,再在此基础上探讨减小抽样框误差的途径。

二、抽样框误差类型及影响

(一) 抽样框误差类型

1. 丢失目标总体单元 这是指抽样框没能覆盖所有总体单元。有些总体单元本属于调查对象,但由于没有在抽样框中出现,因而不可能被选入样本。丢失单元会造成总量估计偏低,也会造成均值(或比例)估计的偏倚。通常,丢失单元的问题不易被察觉,或者即使知道抽样框覆盖不全,但如何寻找丢失单元也很困难。丢失单元是一种威胁性较大的抽样框误差。

2. 包含非目标总体单元 这指的是抽样框中包含了一些本不属于调查对象的非目标总体单元。例如对家庭进行电话调查,在由电话簿组成的抽样框中有一些机构的电话号码,这些机构的号码就属于非目标总体单元。另一种表现是,有些家庭的电话已拆(如家庭搬迁),但原号码仍保留在抽样框中。包含非目标单元使得抽样总体单元个数大于目标总体单元个数,造成总量估计偏高。由于发现非目标总体单元相对容易,并可以通过一定程序将其剔除,所以与丢失目标总体单元相比,包含非目标总体单元的误差影响要小些。

3. 复合连接 这指的是抽样框中的单元与目标总体单元不完全是——对应的关系,一个抽样框单元连接多个目标单元,或一个目标单元连接多个抽样框单元。例如以居住的门牌号作为住户调查的抽样框,一个门牌号内居住两户或多户家庭就属于一个抽样框单元连接多个目标单元的情形,而一户拥有两处或多处住房属于一个目标单元连接多个抽样框单元的情形。复合连接的情况如果严重,将会造成样本的实际抽选与设计要求发生偏离,从而对估计结果产生影响。

4. 不正确的辅助信息。有些抽样设计需要抽样框提供辅助信息,如分层抽样、不等概抽样、比率估计和回归估计等。如果这些辅助信息不完全或不正确,不仅不能提高估计的效率,有时反而会降低估计的准确性。

(二) 对抽样框误差的基本认识

对抽样框误差类型讨论之后,结合实际应用有几点基本认识。

1. 有些误差来自构成抽样框资料的本身。尽管设计工作十分细致,但仍无法避免误差;有些则是因为研究工作不够,资料准备不足,否则有可能建立一个比较好的抽样框。

2. 抽样框中的问题有些容易被发现,有些不容易被发现,即使对于被发现的问题,有些也不容易解决。因此抽样框的维护、抽样框使用情况的不断总结与研讨,对于经常性的调查项目来说是十分必要的。

3. 抽样框不完善并不意味着不能使用。对不完善的抽样框进行修补、调整,有时容易,有时则比较困难,需要一定的财力支持。不完善抽样框是否具有被使用、被

修改的价值,主要取决于抽样框的误差程度、修改后所提高的估计效率、为此所付出的时间和费用以及抽样框的使用次数。

4. 抽样框误差在有些场合会被解释为其他类型的非抽样误差。例如使用地图抽样框,在区域边缘常会出现交错现象,将域内单元划出或将域外单元划入。有些人认为这是抽样框误差,但说成计量误差也有道理。

(三) 抽样框误差的影响

分析抽样框误差影响的一项重要内容,是对抽样框误差造成的偏倚进行定量分析。鉴于丢失目标总体单元是抽样框误差中最常见的一种,故以此为例做稍加深入的分析。

设目标总体由 N_1 个抽样框中单元和 N_0 个抽样框中丢失的单元组成,即 $N = N_1 + N_0$,则总体总和与均值估计的情况如下。

1. 总和估计。总体总和的真值是:

$$Y = \sum_1 Y + \sum_0 Y = Y_1 + Y_0 \quad (11.1)$$

现从抽样框中的 N_1 个单元中采用简单随机抽样抽出容量为 n 的一个样本,由于 n 取自于 N_1 ,为一致不妨记为 n_1 ,对总体总和 Y 的估计为:

$$\hat{Y} = \frac{N_1}{n_1} \sum_1 y_i \quad (11.2)$$

显然此时的估计是有偏的,偏倚为:

$$E(\hat{Y}) - Y = Y_1 - Y = -Y_0 \quad (11.3)$$

这表明估计量低估了总体总和。令

$$r = \frac{Y_0}{Y_1}, W_0 = \frac{N_0}{N}$$

则 Y 的相对偏倚可以写为:

$$-\frac{Y_0}{Y} = -\frac{W_0 r}{r W_0 + (1 - W_0)} \quad (11.4)$$

由上式看出,总体总和估计的相对偏倚取决于 r 和 W_0 两个因素。如果 $r = 1$,即丢失单位均值与抽样框单位均值相同,则相对偏倚为 $-W_0$ 。抽样调查的实践中,抽样框中的丢失单元往往规模较小,一般为 $r < 1$,故相对偏倚的绝对值也就小于 W_0 的绝对值。 r 与 W_0 的关系可从表 11.1 略见一斑。

2. 均值估计。在抽样框存在丢失单元情况下,均值估计量为:

$$\hat{\bar{Y}} = \frac{1}{n_1} \sum_1 y_i \quad (11.5)$$

表 11.1 丢失单元条件下总体总和估计的相对偏倚

丢失单元比重 (W_0)	$r = \frac{Y_0}{Y_1}$				
	0.5	0.9	1.0	1.1	2.0
0.01	0.005 0	0.009 0	0.010	0.010 9	- 0.019 8
0.05	0.025 6	- 0.045 2	0.050	- 0.054 7	0.095 2
0.10	0.052 6	0.090 9	0.100	0.108 9	0.181 8
0.25	0.142 9	0.230 8	0.250	0.268 3	0.400 0
0.50	- 0.333 3	0.473 7	0.500	0.523 8	- 0.666 7

此时估计量的偏倚为:

$$E(\hat{Y}) - Y = W_0(Y_1 - Y_0) \quad (11.6)$$

\hat{Y} 的相对偏倚可以写为:

$$\frac{W_0(Y_1 - Y_0)}{Y} = \frac{W_0(1 - r)}{rW_0 + (1 - W_0)} \quad (11.7)$$

由上式看出,如果丢失单元均值和抽样单元均值相同,即 $r = 1$,则估计量 \hat{Y} 是目标变量 Y 的无偏估计。反之,如果 $r \neq 1$,偏倚状况则随 r 的变化而变化。这种情况见表 11.2。

表 11.2 丢失单元条件下总体均值估计的相对偏倚

丢失单元比重 (W_0)	$r = \frac{Y_0}{Y_1}$				
	0.5	0.9	1.0	1.1	2.0
0.01	0.005 0	0.001 0	0	- 0.000 9	0.009 9
0.05	0.025 6	0.005 0	0	0.004 9	0.047 6
0.10	0.052 6	0.010 1	0	0.009 9	0.090 9
0.25	0.142 9	0.025 6	0	0.024 4	0.200 0
0.50	0.333 3	0.052 6	0	0.047 6	0.333 3

三、不完善抽样框的使用

抽样框不完善并不意味着不能使用,因为构造一个完善的抽样框有时是非常困难的。使用不完善抽样框时若能采用一些补救措施,有助于减小抽样框误差。对不完善抽样框进行补救的具体方法有多种,大致可以分为三种类型。一种是利用核

查或其他有关资料,掌握误差情况,对不完善的抽样框进行调整,或对不完善抽样框所得到的估计量进行调整;第二种是事先制定一些规则,对发现的抽样框问题进行现场处理;第三种是使用多个抽样框进行抽样。下面主要对后两种类型的补救措施做些讨论。

(一) 实行连接

这种方法是事先制定一些规则,使没有包含在抽样框中的目标单元与包含在抽样框中的单元相连接,以弥补抽样框中丢失单元所造成的影响。例如,欲从一份几个月前准备好的学生名单中抽选一个在校学生的样本,新转来学生的名单没有列入抽样框中,因而没有机会入选样本。事先制定的规则为,每个新学生与所在班名单上的最后一名学生相连接,如果最后一名学生被抽中,新学生也就算被抽中并一起接受调查,这样就把可以查明的丢失单元纳入到不完善的抽样框中。这些丢失单元与抽样总体中单元被选中的概率相同,因而得到的估计量也是无偏的。

住户调查中,抽样框可能漏掉一些新建的房屋。如果调查名单上一栋住宅与下一栋住宅的调查路线确定后,那么位于名单住宅之间,而在名单上又漏掉的房屋都可以与刚刚经过的上一栋住宅连接起来。

(二) 惟一连接

抽样框误差的一种类型是复合连接。例如以心血管病患者的就诊病历为抽样框,对心血管病患者进行抽样调查。有些患者在不同的医院看过病,这些人被抽中的概率就高于其他人。可以规定,有两个以上病历者,以最近就医记录的病历号组成抽样框。

(三) 使用多个抽样框

这是指在抽选样本过程中使用两个或多个抽样框。这种方法主要用于抽样框中丢失单元的情况,既然一个抽样框覆盖不全,就采用多个抽样框。在实践中多采用两个抽样框,如名录框和地域框同时使用。使用多个抽样框的主要问题是容易产生重叠现象,如有 A 和 B 两个抽样框,情形如图 11.1。

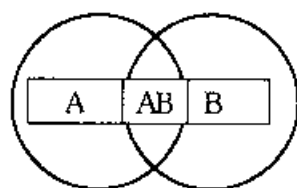


图 11.1 抽样框的重合

图中的 AB 就是重叠部分。重叠会对估计产生影响,解决的办法是剔除重叠。如果抽样框 B 中的单元在抽样框 A 中也存在,就将其剔除。但剔除工作有时十分困

难,甚至无法实施,这就需要利用有重叠的抽样框进行估计。

设样本取自 A、B 两个抽样框,这两个抽样框的单元数分别为 N_A, N_B , 两个抽样框将目标总体划分为三个区域。

区域 a : 其中的单元仅仅与抽样框 A 有联系,单元个数为 N_a ;

区域 b : 其中的单元仅仅与抽样框 B 有联系,单元个数为 N_b ;

区域 ab : 其中的单元与抽样框 A、B 均有联系,单元个数为 N_{ab} 。

现采用简单随机抽样,从 A、B 框中分别抽出容量为 n_A, n_B 的两个样本。利用抽样框 A 的样本对区域 a 和区域 ab 进行事后分层的总和估计分别为:

$$\hat{Y}_A(a) = \frac{N_a}{n_a} y_A(a) \quad (11.8)$$

$$\hat{Y}_A(ab) = \frac{N_{ab}}{n_{ab}} y_A(ab) \quad (11.9)$$

式中, n_a 与 n_{ab} 为落入区域 a 和区域 ab 的样本单元数; $y_A(a)$ 和 $y_A(ab)$ 为区域 a 和区域 ab 的样本观测值总和。

类似地,利用抽样框 B 的样本对区域 b 和区域 ab 进行事后分层的总和估计分别为:

$$\hat{Y}_B(b) = \frac{N_b}{n_b} y_B(b) \quad (11.10)$$

$$\hat{Y}_B(ab) = \frac{N_{ab}}{n_{ab}} y_B(ab) \quad (11.11)$$

于是,目标总体的总和估计为:

$$\hat{Y} = \hat{Y}_A(a) + W_A \hat{Y}_A(ab) + W_B \hat{Y}_B(ab) + \hat{Y}_B(b) \quad (11.12)$$

式中, W_A, W_B 为适当选取的权数,且有 $W_A + W_B = 1$ 。

倘若 n_A, n_B 都足够大,使 $\frac{1}{n_A^2}$ 和 $\frac{1}{n_B^2}$ 可以忽略,且有限总体修正系数 fpc 也忽略不计,则估计量 \hat{Y} 的方差近似为:

$$V(\hat{Y}) \approx \frac{N_A^2}{n_A} [S_a^2(1 - \alpha) + \alpha W_A^2 S_{ab}^2] + \frac{N_B^2}{n_B} [S_b^2(1 - \beta) + \beta W_B^2 S_{ab}^2] \quad (11.13)$$

式中, α, β 分别为重叠部分的单元占抽样框单元的比例,即

$$\alpha = \frac{N_{ab}}{N_A}, \quad \beta = \frac{N_{ab}}{N_B} \quad (11.14)$$

而 S_a^2, S_b^2 和 S_{ab}^2 分别为目标总体三个区域的方差。

确定各个抽样框的样本量 n_A, n_B 和权数 W_A 还需结合调查费用。令

计偏倚。无意无回答可以看成是随机的,这种无回答虽然会造成估计量方差增大,但通常认为不会带来估计偏倚。

当然,如果无回答产生于某个群体,而该群体与其他群体在目标变量方面存在数量差异,那么即便是无意无回答,也会造成估计量的偏倚。例如调查居民的旅游开支,不在家的人可能恰恰是经常外出旅游的。虽然这是无意无回答,但却会造成有偏估计。

二、无回答产生的原因及影响

如果把采集数据的过程划分为查找、接触和采访三个阶段,三个阶段都有可能出现无回答。

1. 查找阶段。调查人员无法找到被调查者,主要原因有地址不详、被调查者搬迁、被调查者不在现场、调查人员不熟悉地址等。

2. 接触阶段。被调查者由于客观原因无法接受调查,如生病或没有时间;被调查者由于主观原因拒访,如对调查不感兴趣,出于安全考虑不让调查员入户等。

3. 采访阶段。调查开始后被调查者对某些问题不愿提供答案、调查员由于粗心遗漏某些项目、由于某种原因调查中断等。

为了分析无回答的影响,可以假设总体由两个层组成。一个是“回答层”,如果这个层的单元被抽中,就可以得到回答;另一个是“无回答层”,采用相同抽样方式,如果这个层的单元被抽中,就无法得到回答。设 N, N_1, N_0 分别为总体单元数、回答层单元数,无回答层单元数。 R_1, R_0 分别为总体回答率和无回答率,即

$$N = N_1 + N_0, R_1 = \frac{N_1}{N}, R_0 = \frac{N_0}{N} \quad (11.19)$$

则总体均值 $Y = R_1 Y_1 + R_0 Y_0$

从总体中抽取容量为 n 的简单随机样本, n_1 来自于回答层, n_0 来自于无回答层。根据回答单元计算出的样本均值为 y_1 ,它是总体中回答层均值的无偏估计,即 $E(y_1) = Y_1$ 。于是用 y_1 作为总体真值 \bar{Y} 的估计值,其偏倚为:

$$\text{偏倚}(y_1) = E(y_1) - Y = \bar{Y}_1 - (R_1 Y_1 + R_0 Y_0) = R_0(Y_1 - Y_0) \quad (11.20)$$

$$\text{相对偏倚}(y_1) = \frac{R_0(Y_1 - \bar{Y}_0)}{Y} \quad (11.21)$$

相同的方法可以得到总量估计的偏倚和相对偏倚分别为:

$$\text{偏倚}(\hat{y}_1) = NE(\hat{y}_1) = NY - NR_0(Y_1 - Y_0) \quad (11.22)$$

$$\hat{R} \text{ 相对偏倚}(\hat{y}_1) = \frac{NR_0(Y_1 - Y_0)}{NY} = \frac{R_0(Y_1 - Y_0)}{Y} \quad (11.23)$$

这表明,总量估计的绝对偏倚等于均值估计的绝对偏倚乘以总体单位数 N ,总量估计和均值估计的相对偏倚相等

由模型看出,导致无回答偏倚的原因主要来自于两个方面:一个是回答层与无回答层单位之间的数量差异($Y_1 - Y_0$);一个是无回答率 R_0 。

上述模型给我们一些启示:首先,如果 $Y_1 = Y_0$,即回答单元与无回答单元目标变量的数量特征没有显著差异,可以看成无回答是由于一些随机因素引起的,这时问题尚不严重,因为不会引起估计偏倚。但是,由于无回答造成实际接受调查单元数目减少,会引起估计方差的增大,这时只要简单地增大样本量,使完成调查单元数目与调查方案要求相一致即可。其次,如果 $Y_1 \neq Y_0$,仅仅用回答数据进行估计就会产生偏倚,且 Y_1 与 Y_0 差异越大,估计偏倚就越大,这时降低无回答率就十分重要。最后,如果无法最终消灭无回答,就需要采取一些补救措施,通过对调查数据的调整,以减小由于无回答对估计带来的影响。

三、降低无回答的措施

解决问题的最好方法是在问题发生之前采取措施加以预防,对调查中的无回答也是如此。导致无回答的原因是多方面的,如果调查进行前对可能产生无回答的原因加以认真研究,并有针对性地采取预防措施,就会有效地提高调查中的回答率。

可以采用的预防措施主要有:

1. 问卷设计具有吸引力,容易引起被调查者参与的兴趣,并注意适当的长度。
2. 在可能的条件下,充分利用调查组织者的权威性和社会影响力,激发被调查者的参与意识。
3. 确定准确的调查方位,使调查员容易找到被调查者。
4. 采取有助于消除被调查者冷漠、担心或怀疑的措施,如预先通知、调查前的解释说明及雇用与被调查者熟悉的人做调查员。
5. 注意调查员的挑选。调查员的身份与被调查者越接近,就越容易被对方接受。实践表明,大学生、居民委员会成员、下岗职工都是理想的非专职调查员人选。
6. 做好调查员的培训,增强调查员的责任心,提高其访谈技巧。有经验的调查人员可以把调查中的无回答率降到最低程度。
7. 注意调查过程的监控。对不成功的调查及时总结,找出解决问题的办法。如拒访是什么原因造成的,调查时间是否合适。要让一个球迷在一场精彩的球赛转播时接受调查,其难度可想而知。

8. 奖励措施。调查总要花费被调查者的时间和精力,适当的奖励是必要的。如邮寄调查中采用抽奖,入户调查中向被调查者赠送小礼品,对集体单位进行调查时许诺提供最后的调查报告或汇总结果,等等。一些人接受调查并不是为了得到奖励,但奖励措施会使对方感到他们提供的信息是多么重要。

9. 再次调查。再次调查是指在概率抽样的第一轮调查完成之后,针对无回答产生的原因,采取相应的措施,对无回答单元进行再次的调查。无回答产生的原因包括:

(1) 不在家。调查人员了解到调查对象何时在家,再次登门调查。

(2) 不方便。调查时被调查者由于生病、工作忙或其他客观原因难以接受调查,调查人员可以约定另外的时间,在对方方便的时候进行调查。

在上述两种情况下,再次调查都可以收到明显效果。此外,对一些不明原因的拒访,可以改变调查方式。例如,对于邮寄调查的无回答者,除再次邮寄调查问卷外,可以用电话提醒或改用电话调查。作为一般的原则,应该对被调查者尝试三次,若仍不成功,才可将其放弃。

10. 替换被调查单元。对于放弃的无回答者,需要抽取替换单元,以便使接受调查的样本单元数不低于原设计要求。替换的原则应该事先规定,例如入户调查中的“右手原则”,即用放弃户右边的第一户作为替代单元。替代原则的事先规定可以防止调查员自作主张,也便于事后检查。

影响回答率的一大障碍是调查中的敏感问题,所以调查问卷中应尽量避免敏感性问题。但有些调查本身就是针对敏感问题的,由此提出了随机化回答技术。随机化回答技术的基本特征是,被调查者对所调查的问题采用随机回答的方式,从而对自己的回答起到匿名的作用。调查人员根据事先设计的程序,可以对目标变量进行推算。针对不同的目标变量,有不同的随机化回答模型。

四、对存在无回答数据的调整

调查中无回答的情况总是难以避免。由于无回答造成数据不全,如果不加处理,就有可能造成估计量偏倚。针对不同的情况,可以考虑采用一些补救措施,以对无回答造成的估计量偏倚起到纠偏、校正的作用。对存在无回答数据进行调整的方法有多种,下面介绍其中的几种。

(一) 再抽样调整

这种方法是指在第一次无回答的单元中随机抽取一个子样本,通过更细致、更充分的工作,获得该子样本的数据,作为整个无回答层的代表值。然后把第一次调查中回答层数据和第二次调查中无回答层数据结合起来,对总体的有关参数进行

估计. 设从总体 N 中随机抽取 n 个样本单元, 第一次调查有 n_1 个回答单元和 n_0 个无回答单元, $n = n_1 + n_0$; 再从 n_0 个无回答单元中随机抽取一个容量为 m 的子样本进行调查, 令 y_1 和 y_0 分别代表第一次 n_1 个单元和第二次 m 个单元的样本均值, 则可以得到总体均值 Y 的无偏估计:

$$\hat{Y} = \frac{1}{n}(n_1 y_1 + n_0 y_0) = u_1 y_1 + u_0 y_0 \quad (11.24)$$

式中, $u_1 = \frac{n_1}{n}$, $u_0 = \frac{n_0}{n}$ 分别为样本中回答层和无回答层的比例

抽样用到两个随机程序: 一次是从 N 个单元中随机抽取 n_1 个单元; 另一个是从第一次无回答的 n_0 个单元中随机抽取 m 个单元. 根据抽样估计原理, 目标变量经过两个随机程序的方差是:

$$V(\hat{Y}) = V_1 E_0(\hat{Y}) + E_1 V_0(\hat{Y}) \quad (11.25)$$

第二个随机程序的条件期望值和估计量方差分别是:

$$E_0(\hat{Y}) = \frac{1}{n}(n_1 y_1 + n_0 y_0) = y \quad (11.26)$$

$$V_0(\hat{Y}) = u_0^2 \frac{(k-1)}{n_0} s_0^2 = w_0^2 \frac{(k-1)}{n} s_0^2 \quad (11.27)$$

式中, s_0^2 为样本无回答层的方差; k 为无回答层抽样间距, 即 $k = \frac{n_0}{m}$.

将式(11.26)、式(11.27)代入式(11.25), 便有

$$\begin{aligned} V(\hat{Y}) &= V_1(y) + E_1 \left[u_0^2 \frac{(k-1)}{n} s_0^2 \right] \\ &= \frac{1}{n} f S^2 + w_0^2 \frac{(k-1)}{n} S_0^2 \end{aligned} \quad (11.28)$$

式中, S^2 为总体方差; S_0^2 为总体中无回答层的方差

上式等号右边的第一项是通常情况下的简单随机抽样误差计算公式, 第二项是采用再抽样后方差的增加部分。可以看出, 当无回答层所占比例 w_0 较小时, 进行再抽样所增加的估计量方差就比较小, 特别是当 $k = 1$, 即对样本中所有的无回答都进行再次调查并获得回答时, 第二项方差部分为零, 这时的估计量方差就等同于样本量为 n 的简单随机抽样。对无回答层单元进行再抽样, 把两次调查的数据结合起来, 可以得到目标量的无偏估计, 从而实现了“对缺失数据可能带来估计偏倚进行校正的目的”, 但它是“以增大估计量方差为代价的”。

(二) 加权调整

对存在无回答数据进行补救的另一种方法是采用加权调整。加权调整法是通

过对调查中所获得的回答数据使用加权因子,达到对数据进行调整,减小由于无回答造成的估计偏倚。作为说明,设从总体 N 中随机抽取容量为 n 的样本,估计量

$\hat{Y} = \sum_{i=1}^n W_i Y_i$ 是无偏的,这里 W_i 是第 i 个样本单元的权数;若令 π_i 为第 i 个单元的入样概率,在样本单元全部回答情况下,权数 $W_i = \pi_i^{-1}$,它反映了第 i 个样本单元在估计中的作用。又设 P_i 为第 i 个单元的回答概率, $P_i = 1$ 表示一定回答, $P_i = 0$ 表示一定不回答,现实中 P_i 是一个随机变量,被调查者是否回答取决于多种因素。设回答概率期望值 $E(P_i | \pi_i = 1) = P$,即第 i 个单元被选中后的回答概率为 P 。在调查中,由于无回答的存在,只能用 n_i 个回答单元的信息对总体参数进行估计,因此

此估计量 $\hat{Y} = \sum_{i=1}^n W_i Y_i$ 就需要修正为 $\hat{Y}^* = \sum_{i=1}^n W_i^* Y_i$,其中 $W_i^* = (\pi_i P_i)^{-1}$ 是对无回答数据进行调整的权数。从这个意义上说,调整是根据调查中回答单元的回答概率进行的。

为进行调整,需要掌握样本单元的回答概率。由于 P_i 未知,就需要对 P 进行合理的估计,对 P 的不同估计就形成不同的调整方法。因此,加权调整法是一个概括的说法,它包括了一些不同的调整方法。这里介绍最基本的加权组调整 (weighting class adjustment) 方法。

首先,将容量为 n 的样本划分为 H 个互不重叠的子集,把这些子集称为调整组,用下标 h 表示。通过划分使得组内各单元的目标变量 Y_i 值尽可能相近,并假设组内所有单元的回答概率相同。这个过程类似于对样本进行分层,因而需要足够的进行分层的辅助信息。

加权组调整中使用的 P_i 的估计量,是组内经过加权的回答率。依前述,对任何概率样本,有 $W_{hi} = \pi_{hi}^{-1}$,这里 W_{hi} 是第 h 组中第 i 个样本单元的未经调整的权数, P_{hi} 的估计量为:

$$\hat{P}_{hi}^{(2)} = \frac{\sum_{i=1}^{n_h} W_{hi} Y_i}{\sum_{i=1}^{n_h} W_{hi}} \quad (11.29)$$

式中, n_h 为第 h 组中的样本量; n_{1h} 为第 h 组中回答单元的个数; $\hat{P}_{hi}^{(2)}$ 为第 h 组第 i 个单元回答概率的估计值。

于是经过加权组调整的权数为:

$$W_{hi}^{(2)} = \frac{W_{hi}}{\hat{P}_k^{(2)}} = W_{hi} \cdot \frac{\sum_{i=1}^{n_h} W_{hi}}{\sum_{i=1}^{n_h} W_{hi}} \quad (11.30)$$

显然,如果没有无回答, $\hat{P}_k^{(2)} = 1$, 调整后和调整前的权数没有什么区别. 如果存在无回答, $\hat{P}_k^{(2)} < 1$, 则 $W_{hi}^{(2)} > W_{hi}$, 它表明, 由于无回答单元无法提供信息, 有关无回答单元的信息被分摊到回答单元的身上. 还可以看出, 如果采用等概率抽样设计, 即 $\pi_{hi} = \frac{n}{N}$ 对所有的 $h = 1, 2, \dots, H$ 都成立, 则 $P_{hi}^{(2)} = \frac{n_{1h}}{n_h}$.

令 $\Delta_h = \frac{N_h}{N}$ 是第 h 组在总体中所占的比重, 则 Δ_h 的估计值为:

$$\hat{\Delta}_h = \frac{\sum_{i=1}^{n_h} W_{hi}}{\sum_{h=1}^H \sum_{i=1}^{n_h} W_{hi}} \quad (11.31)$$

令 $Y_{1h} = \frac{\sum_{i=1}^{n_{1h}} Y_{hi}}{N_{h1}}$ 为第 h 组中回答层的总体均值, 则 Y_{1h} 的估计值为:

$$\hat{Y}_{1h} = \frac{\sum_{i=1}^{n_{1h}} W_{hi} \cdot Y_{hi}}{\sum_{i=1}^{n_{1h}} W_{hi}} \quad (11.32)$$

由式(11.30)、式(11.31)、式(11.32), 可以得到总体均值估计为:

$$\hat{Y}_{\text{adj}} = \sum_{h=1}^H \hat{\Delta}_h \hat{Y}_{1h} = \frac{\sum_{h=1}^H \sum_{i=1}^{n_{1h}} W_{hi}^{(2)} Y_{hi}}{\sum_{h=1}^H \sum_{i=1}^{n_{1h}} W_{hi}^{(2)}} \quad (11.33)$$

估计量下标 adj 表示加权组调整.

(三) 相关推估法

相关推估法主要用于调查中的项目无回答. 项目无回答指被调查单元不是完全拒绝调查, 而是拒绝其中某些项目的调查. 这时其他回答项目的信息尚可以利用, 利用这些信息, 对无回答的数据进行推估. 基本思路是, 寻找与无回答问题变量有关联的其他调查问题变量, 利用调查数据建立起变量之间的回归方程, 对项目无

回答的变量值进行推估。例如,如果我们认为居住面积(当然,还可能有其他项目,如职业、职位,拥有高档耐用品种类和数量等)与收入有关,被调查者的居住面积是可知的,就可以建立起收入与居住面积的回归方程,如果方程拟合效果好,就可以对收入项的无回答进行推估。现场调查中,除了无回答以外,还会有一些其他原因造成缺失数据,如遗漏、丢失,或在数据审核中将明显的不合逻辑的数据删除等。对缺失数据进行推估,除了回归法以外,也还有其他许多方法。例如,某企业利润数明显不实,将其剔除,在同次调查中得到的其他有关数据如表 11.3 所示。

表 11.3 相关推估法示例

项 目	某企业	同行业同规模 其他企业平均值
销售量(箱)	1 000	700
利润(万元)	—	15

由于利润与销售量有很高的相关度,用该行业相同规模的其他企业的调查结果可以推估某企业的利润约为 21.4 万元($15 \times \frac{1\,000}{700} = 21.43 \approx 21.4$)。

(四) 插补调整

“插补”一词译自于 imputation,该词又有估计、推算、替代等多种译法。其基本意思是,在数据整理阶段,利用调查结果,采用一定的方式,为无回答的缺失值确定一个合理的估计值,插补到原缺失数据的位置上。插补可以达到两个调整目的:一是减小由于无回答可能造成的估计量偏倚,为此,就要使确定的替补值尽可能地接近缺失的原数据值。事实上缺失数据的真值人们无法得知,因此所追求的只能是确定替补值方法的合理、有效。调整的第二个目的是力图构造一个完整的数据集。在调整前,由于无回答的存在,使原数据集上出现许多“窟窿”,给一些统计分析方法的使用带来不便。采用插补的方式填补了缺失值的空缺,就为后面分析人员的工作提供了方便,他们在使用标准统计软件的同时,不必烦琐地说明对缺失值进行处理的方法,大大节省了精力和时间。而且不同分析人员使用的是同一套经过插补调整的数据,也保证了分析结果的一致性。插补的效率如何,取决于替补值与缺失值的近似程度。为了提高效率,对研究总体进行分层,使层内各单元诸方面情况尽可能相似,利用同一层内回答单元的信息产生出缺失数据的替补值是进行插补的基本思路。因为可以利用不同的信息源,采用不同的方式生成替补值,所以有不同的插补方法。

实际中使用较多的是均值插补,其方法为:首先根据辅助信息将样本分为若干组,使组内各单元的主要特征相似。然后分别计算各组目标变量 Y 的均值,将各组

均值作为组内所有缺失项的替补值。均值插补法的特点是操作简便,并且对均值和总量这样的单变量参数可以有效地降低其点估计的偏倚。但它的弱点也比较突出,首先是插补的结果歪曲了样本单元中 Y 变量的分布,因为同组中无回答的替补值都由该组的平均值充当,使得其分布状况受到由各组回答单元数据计算出的组均值的制约;其次,插补结果将导致在均值和总量估计中对方差的低估,因为同一组内样本单元的离差将由于同一个数值的多次出现而偏低,因此均值插补适用的场合是仅仅进行简单的点估计,而不适用于需要方差估计等比较复杂的分析。

为避免均值插补中替补值过于凝集的弱点,可以使用随机插补,这种方法是指采用某种概率抽样的方式,从回答单元的资料中抽取无回答的替补值。为便于说明,令某项目回答数据个数为 n_1 ,无回答个数为 n_0 ,则 $n = n_1 + n_0$,现从 n_1 个数据中随机抽取 n_0 个替补值,则样本构成为:

$$\text{样本} = y_1, y_2, \dots, y_{n_1}, y_{n_1+1}^*, y_{n_1+2}^*, \dots, y_n^*$$

此时,目标变量的均值估计为:

$$\bar{y} = \frac{1}{n} (n_1 \bar{y}_1 + n_0 y^*) \quad (11.34)$$

$$\text{式中, } y^* = \sum_{i=1}^{n_0} \frac{H_i y_i}{n_0} \quad (11.35)$$

如果采用不重复抽样, $H_i = 0$ 或 1 ; 如果采用重复抽样,则 H 为多项式分配,若 $h_1 + \dots + h_{n_1} = n - n_1$, 则

$$P[H = (h_1, h_2, \dots, h_{n_1})] = \frac{(n - n_1)!}{h_1! h_2! \dots h_{n_1}!} \quad (11.36)$$

否则上面所定义的概率为零,由此得出

$$E(H_i) = \frac{n - n_1}{n_1} \quad (11.37)$$

$$\text{Var}(H_i) = (n - n_1) \left(1 - \frac{1}{n_1}\right) \frac{1}{n_1} \quad (11.38)$$

$$\text{Cov}(H_i, H_j) = -\frac{n - n_1}{n_1^2}, i \neq j \quad (11.39)$$

若假设 i 与 j 独立,由上面结果可以导出

$$E(y) = Y \quad (11.40)$$

$$\text{Var}(y) = \left(\frac{1}{n_1} - \frac{1}{N}\right) S^2 + \left(1 - \frac{1}{n_1}\right) \left(1 - \frac{n_1}{n}\right) S^2 \quad (11.41)$$

式中, S^2 为总体方差。

可以看出,随机插补法估计量 y 的方差由两部分组成,等式右边第一项是仅用

(二) 被调查者误差

在计量误差中,被调查者误差是最重要的,这是指被调查者在调查过程中没有给出真实的回答。当然这里不包括前面提到的由于对问题不理解等原因造成的误差,可以把被调查者误差划分为两类:无意识误差和有意识误差。无意识误差的主要表现是回忆误差,指对调查内容记忆不清而回答失真,无意识误差还包括“倾向性数字”心理学研究表明,人们在回答数字问题时,常常会下意识地给出一些倾向性数字。如调查吸烟者每天的吸烟量,回答往往是一包、半包或10支;调查人们每天看电视的时间,回答经常是半个小时、1个小时等,尽管真实的数值并非如此。被调查者有意识的回答误差则多是由于问题的敏感或其他因素使回答具有某种倾向性,如回答学历、职务职称时,往往有高报倾向,而对另一些调查内容,由于利益驱使往往低报。无意识误差可以看做是随机的,不会带来估计偏倚,有意识误差则不然,由于它存在倾向性,所以会导致严重的估计偏倚。一般而言,这种误差的倾向性根据经验和实际情况的分析是可以察觉的。

(三) 调查者误差

这是指由于调查员的原因而引起的误差。有些是由于调查员工作不认真,如记录错误等造成的,有些则是调查员在调查过程中,将自己的思想、观点、看法、感受等与被调查者交流,对被调查者的回答起了“诱导”作用。

(四) 其他误差

计量误差的产生还有其他一些来源,这里列举一些。

1. 测量工具 在需要利用测量工具进行的调查中(如农产量抽样调查的采样框和磅秤,学生视力调查中与测量表的距离和光线等),如果测量工具不准,就会对测量结果带来偏误,即便测量工具是精确的,反复测量也会产生随机误差。

2. 编码 编码的作用是把数据变为可机读的形式。编码错误不仅仅指具体的编写错误,也包括对编码结果的理解上。特别是对于调查中开放性问题的编码,同样的内容,由于理解不同,不同编码员的编码结果可能不同。

3. 录入 要求数据录入过程中不发生任何错误是很难做到的,只能采取一些措施,把录入错误降到最低限度,如使用双机录入等。

综上所述,计量误差内容繁杂,它对于调查数据质量的影响是不可忽视的。

二、计量误差模型

国内外已有大量的文献对计量误差模型展开讨论,这里仅对计量误差最基本的模型做些分析,使读者对此有所了解。

在理论上可以假设对第 i 个单元进行多次重复性调查并做计量,令

$$y_{it} = \mu_i + e_{it} \quad (11.42)$$

式中, μ_i 为第 i 个单元真值; e_{it} 为第 i 个单元第 t 次计量中的误差。

关于 μ_i 的内涵需要做些说明。有些情况下 μ_i 是具体存在的一个确定值。例如, 在一定时点上人的身高、体重, 某职员上个月的收入, 等等。有些情况下, μ_i 又是抽象模糊的, 很难定义其真值。例如采用量表方式对人们的态度、看法、情感等抽象内容的调查, 其计量结果与当时的环境、气氛及被调查者心情关系很大。尽管如此, 进行分析时这个概念仍是不可缺少的。

在对同一个单元进行重复计量情况下, e_{it} 将遵从一概率分布。通常假定是正态分布, 并令

$$E(e_{it}) = B_i \quad (11.43)$$

式中, B_i 为计量中的偏倚。如果 $B_i = 0$, 说明虽有计量误差, 但它是随机的, 其期望值为零; 反之, 若 $B_i \neq 0$, 则表明对 i 单元的计量中存在系统性偏倚。

对于特定的 i 单元, 偏倚 B_i 是个常量, 但对于不同的 i , B_i 可能不同, 若

$$E(B_i) = B \quad (11.44)$$

则称 B 为所有单元的常数偏倚

不妨令

$$d_{it} = e_{it} - B_i \quad (11.45)$$

式中, d_{it} 为对每个单元 i 在第 t 次计量时的误差波动部分。显然, d_{it} 与 e_{it} 有相同的分布, 其期望值 $E(d_{it} | i) = 0$ 。因此, 式(11.42) 又可用

$$y_{it} = \mu_i + e_{it} = \mu_i + B_i + d_{it} \quad (11.46)$$

表明单元 i 的具体观测结果, 它受其均值、计量系统偏倚及计量随机误差几个因素的影响。

进一步令

$$\mu'_i = \frac{1}{t} \sum_t y_{it} \quad (11.47)$$

是对单元 i 进行 t 次计量后的平均, 也即

$$\mu'_i = E(y_{it} | i) = \mu_i + B_i \quad (11.48)$$

则

$$d_{it} = y_{it} - \mu'_i \quad (11.49)$$

如前所述, d_{it} 是计量过程中的随机误差, 它所表现的是实际测量值 y_{it} 与包括偏倚在内的测量均值 μ'_i 之间的差异

由式(11.49), 有

$$y_i - \mu = \mu'_i - \mu + d_i \\ = d_i + (\mu'_i - \mu') + (\mu' - \mu) \quad (11.50)$$

式中,

$$\mu' = \frac{1}{N} \sum \mu'_i \quad (11.51)$$

是 μ'_i 在总体中的均值。

在调查中,抽取容量为 n 的样本,令

$$\mu = \frac{1}{n} \sum \mu_i \quad (11.52)$$

对样本加以平均,则可以将式(11.50) 写为:

$$y_i - \bar{u} = d_i + (u' - u') + (u' - u) \quad (11.53)$$

式中, \bar{u} 为真值 u_i 的均值。由此得到均方误差的公式:

$$MSE(y_i) = V(d_i) + V(u') + (u' - u)^2 + 2\text{cov}(d_i, u') \quad (11.54)$$

等式右边的第一项为计量随机误差,第二项为抽样方差,第三项为计量偏倚的平方,最后一项为协方差,由于有 $E(d_i \setminus i) = 0$,故此项通常为零。

式(11.54)说明了以下几个问题:

第一,如果计量中存在偏倚,结果会使估计量产生偏倚。但依据样本资料无法计算偏倚,因为真值 u 未知。对计量过程中的偏倚识别,需要利用其他有关资料,在某种程度上更需要调研人员的经验以及对调查对象的了解。在可能条件下,通过努力,在小范围内获取被调查单元真值,借以对偏倚进行推算。

第二,偏倚虽然可以影响估计量,但不会影响方差估计。因为如果每个 y_i 中都包含偏倚,其均值 y 中也包含偏倚,在计算 $\sum (y_i - y)^2$ 过程中,偏倚部分相互抵消。

第三,假如不存在常数偏倚,且样本中计量误差 d_{ii} 互不相关,便会有

$$V(y_i) = V(d_i) + V(\mu) \\ = \frac{1}{n} \sigma_d^2 + \frac{1}{n} f S_u^2 \quad (11.55)$$

即使采用全面调查, $f = 1$, 等式右边后一项的抽样方差不再存在,但计量方差仍是存在的。在抽样调查中,由于计量方差的存在,会使总的方差增大,所以一般的抽样误差计算公式往往低估了实际中的误差状况。

第四,若计量误差 d_{ii} 之间存在相关,如在同一个地点接受视力检测的人员受到相同检测条件的影响;若干名在同一区域进行调查的工作人员,接受的是同一位指挥者的技术培训,计量误差之间的相关就是可能的。这时

$$V(d_i) = \frac{1}{n} \sigma_u^2 [1 + (n-1)\rho_u] \quad (11.56)$$

式中, ρ_u 为样本内相关系数。即使 ρ_u 很小, 对 $V(d_i)$ 也会产生极大影响。例如, 如果 $\rho_u = 0.1$, $n = 100$, 则 $1 + (100-1)0.1 = 10.9$, 即 $V(d_i)$ 为原来的 10.9 倍。这说明调查实施中工作人员的规范操作是多么重要。

第五, 将计量方差 $V(d_i)$, 抽样方差 $V(u)$, 偏倚平方 $(u' - u)^2 = B^2$ 用另一种方式表示, 则均方误差公式又可写为:

$$MSE(y_i) = \frac{1}{n} S_u^2 + \sigma_u^2 [1 + (n-1)\rho_u] + B^2 \quad (11.57)$$

可以看出, 随样本量 n 的增大, 抽样方差 $\frac{1}{n} S_u^2$ 会越来越小, 但偏倚平方 B^2 与 n 无关; 在 $\rho_u \neq 0$ 条件下, $(n-1)\rho_u$ 反而会增大。也就是说, 在大样本调查中, B 与 ρ_u 所带来的影响成为均方误差中的主要部分, 抽样方差在总误差中反而显得不太重要。清醒地认识这一点有助于我们认识调查过程中质量控制的重要性。

三、减少计量误差的措施

计量误差涉及的内容广泛, 减少计量误差需要对调查全过程进行质量控制。

(一) 调查设计方面

调查设计的质量与设计人员的能力密切相关。有能力的调查人员能够设计出更好的调查问卷和抽样程序, 以减少由于设计不周所可能带来的计量误差。调查问卷设计出来后, 应组织有关人员对问卷进行讨论。如果是大规模的调查活动, 还应在正式调查之前进行预调查, 在实践中对问卷进行检验。调查设计是整个调查活动的起点, 其专业技术性较强, 对人员素质和技能的要求很高, 一旦设计出现问题, 损失往往是难以补救的。如果设计人员具有丰富的专业知识, 又了解实际情况, 由调查设计所引起的误差是可以得到有效预防的。

(二) 现场准备方面

在收集数据之前, 需要做许多准备工作, 这些工作质量的好坏, 对计量误差会产生直接影响。主要的准备工作包括招聘访问员、对访问员培训、编写调查手册。

1. 招聘调查员。每一个调查机构, 通常都会有一份访问员名单, 名单上记载的是经过培训的访问员, 包括固定员工和以前调查所雇用过的访问员。调查机构可以根据这份名单, 招聘调查所需的访问员。但是如果调查需要大量访问员, 就需要招聘新的访问员。在任何情况下, 调查所需要访问员的条件都应该明确。招聘访问员时, 其文化程度、沟通能力、语言能力、组织能力和思想素质都是应考虑的重要因素。如果进行大范围电话调查, 由于区域跨度大, 招聘访问员时还应考虑用不同

地区方言进行交流的问题。

2. 培训访问员。实践证明,访问员的培训对调查数据质量起着近乎决定性的影响。培训内容通常有调查内容的培训(熟悉调查问卷和调查工作程序)和调查技能的培训(如何处理调查过程中遇到的疑难问题)培训方式有课堂讲授、模拟面访和实习面访等。在培训过程中,能否充分调动访问员的工作热情,帮助访问员树立克服各种困难的坚定信念和决心,是衡量培训成功与否的一个重要标志。

3. 编写调查手册。调查手册是访问员进行工作的指南。好的调查手册有助于访问员更有效地开展工作。调查手册的内容通常包括:调查内容(调查问卷)的说明,问卷的审核规则,作业管理(如怎样报告调查进程,怎样分发和回收问卷,调查所需的设备和材料等)的规定,以及访问技巧和技术的介绍。

(三) 调查结果审核方面

审核是对调查质量进行控制的一道工序,也是减少计量误差的有效方法。审核的目的是要保证调查所得到数据的完整性、一致性和有效性。审核工作贯穿于整个调查过程。

审核有三种类型,即有效性审核、一致性审核和数据分布审核。有效性审核是检查调查数据是否有效,包括是否在需要填写数字的地方填上了非数字字符,编码数据是否在允许值之内等。一致性审核主要检查不同问题之间的关系是否正确,它可以基于不同问题或同一问题的不同部分之间的结构关系、逻辑关系来进行。例如,出生年月和婚姻状况,对于22周岁以下的男性公民或20周岁以下的女性公民,婚姻状况除了“未婚”之外,不可能有别的选择;又如,如果问题A回答“否”,问题B就不用回答;等等。数据分布审核通过拟和数据的分布,确认异常记录,然后采取相应的处理方法(如重新核实或删除)。

审核可以在调查过程中的任何阶段进行。

1. 收集数据时进行审核。收集数据时可以做现场审核。访问员在调查进行过程中根据常识或经验,可以判断出一些问题的答案是否属于“可接受”范围。在调查结束后,立即审核所做的记录,由于刚才的信息还记忆犹新,很容易找到被调查者并查明确切情况,因此,有机会发现并纠正错误。

2. 数据收集完毕后的审核。通常,比较全面、比较复杂的审核是在数据收集完毕后进行的。可以把审核视为一个独立的工作环节。审核工作可以由了解情况、经验丰富的专门审核人员进行,也可以由计算机的审核程序来执行。计算机硬件和软件的发展使得进行自动化审核越来越成为可能。在这个阶段,虽然也进行数据有效性的审核,但侧重点是数据的一致性审核和离群值的检测。

$$d = \frac{v - m}{s} \quad (11.58)$$

如果 d 超出了预先确定的偏离值,那么该观测值就被认为是离群值。

另外,离群值也可以通过下面的置信区间进行确认:

$$(m - t_{\alpha} s, m + t_{\alpha} s) \quad (11.59)$$

式中, t_{α} 和 $t_{1-\alpha}$ 分别为根据预先确定的置信度得到的标准正态分布下限和上限的值。如果总体是偏态的, t_{α} 和 $t_{1-\alpha}$ 就要用不等的值,落在这个区间之外的观测值被认为是离群值。

样本均值和样本方差是用来测度数据集中趋势和离散趋势最常用的统计量。但由于它们对离群值比较敏感,因此选择它们就不太合适。例如,如果数据呈偏态分布,样本均值就会偏向离群值,样本方差也会由于离群值而放大。因此,有些离群值的 d 值就会显得相当小,确认这些离群值就较为困难,这种现象称为屏蔽效应。

因此,最流行的检测方法之一是使用四分位数法。这种方法用中位数测度数据的集中趋势,四分位域测度数据的离散程度,因为这些统计量对离群值不太敏感(中位数和四分位数是用加权的样本数据计算出来的)。四分位数把数据分成四个部分:25% 的数据小于第一个四分位数 $q_{.25}$, 50% 的数据小于第二个四分位数(或中位数) $q_{.5}$, 75% 的数据小于第三个四分位数 $q_{.75}$ 。

上、下四分位域 h 和 h_{α} 定义如下:

$$h_{\alpha} = q_{.5} - q_{.25} \quad (11.60)$$

$$h_{\alpha} = q_{.75} - q_{.5} \quad (11.61)$$

置信区间为:

$$(q_{.5} - t_{\alpha} h_{\alpha}, q_{.5} + t_{\alpha} h_{\alpha}) \quad (11.62)$$

其中, t_{α} 和 $t_{1-\alpha}$ 可以通过检查以前的数据或基于过去的经验来确定。任何落到这个区间之外的观测值都被认为是一个离群值。

关于离群值检测方法的详细内容,请参见 Barnett and Lewis (1995)。

三、离群值的处理

对于在调查过程中发现的离群值,可以用几种方法来处理。如果在调查进行中发现离群值,就要及时处理,例如进行回访核实,对错误进行更正。如果在调查完毕后的审核中发现离群值,回访核实已不可能,通常对离群值采用插补处理,即将离群值剔除,然后使用插补法调整。有些情况下,如果认为离群值无大碍,也可以对离群值不做任何处理。这时,主观判断就非常重要,因为忽略或纠正离群值对数据的质量有较大影响。

对在审核时没有进行处理的离群值可以在估计的时候处理。忽略未处理的离群值会影响估计的效果,使估计结果产生偏倚,并导致估计量的方差增大。处理的目的是要引入较大偏倚的前提下,减少离群值对估计量抽样误差的影响。

估计时有三种方法可以处理离群值:(1) 改变数值;(2) 调整权重;(3) 进行稳健估计(robust estimation)。

如果离群值的出现是由某些变量的极值导致的,应该用改变数值或进行稳健估计的方法处理;如果离群值的权重很大,即影响大的离群值,则应该考虑修改其权重,用一种客观的估计方法来减轻它的影响。

(一) 改变数值

处理离群值的一种方法是缩尾化。这种方法首先要将样本数据按从大到小依次排序,然后再按下面的步骤计算。

在简单随机抽样中,总体总量 Y 的无偏估计公式为:

$$\hat{Y} = \frac{N}{n} \sum_{i \in s} y_i$$

式中, i 为样本中第 i 个单元; s 为所有样本单元的集合(假定回答率为 100%)。

类似地,对于缩尾化,假设 $y_i (i = 1, 2, \dots, n)$ 是一系列有序样本数据,来自大小为 N 的一个总体,样本量为 n 。若样本数据中第 k 个最大值 $y_{(n-k+1)}$ 被认为是离群值,单侧 k 次缩尾估计量就可以通过用第 $n-k$ 个最大的值 $y_{(n-k)}$ 代替这些离群值来定义,即

$$\hat{Y}_w = \frac{N}{n} \left(\sum_{i=1}^{n-k} y_i + k y_{(n-k)} \right) \quad (11.63)$$

需要提及,缩尾化适合于处理单个变量的情况,因此它在多变量的抽样调查中很少应用。

(二) 调整权重

处理离群值的另一种方法是降低离群值的权重,从而使它们的影响变小,例如,赋予离群值的权重为 1,即离群值仅代表它自己而不代表其他总体单元。但这样做对估计的影响很大,特别是对偏态总体的估计结果通常为低估。例如,如果某一行业中两个大公司的零售额占总行业零售额的大部分,其中一个公司被选入样本,其权数为 2,因为它代表两个规模类似的单元。但如果这个公司的零售额被确定为离群值,并改变其估计权数,就会严重低估整个行业的总零售额。目前,专家们已经提出了一些能够降低离群值权重的估计量,参见 Rao (1970), Hidiroglou and Srivastava (1981)。

(三) 选取稳健估计量

经典的估计理论中,总体参数的估计量基于某种分布的假设。通常,假定估计

量服从正态分布,样本均值和样本方差估计量在正态分布的假设下也是最理想的。但是,这些估计量对离群值非常敏感。稳健估计量则能克服这种局限性,因为它对分布的假设不太敏感。比如,中位数比均值更稳定;四分位域比通常的方差估计量更稳定。过去几年中,已经提出了很多稳健(robust)估计量。

关于稳健估计量和离群值检测的详细讲解,请参见 Barnett and Lewis (1995), Rousseeuw and Leroy (1987), Lee et al (1992), 以及 Lee (1995)。

小 结

本章对非抽样误差的产生来源和处理方法进行了一般性的讨论。非抽样误差是影响统计调查数据质量的重要方面,在很多情况下,非抽样误差已经超过甚至大大超过抽样误差,对此应引起重视并加以认真研究。本章分别对抽样框误差、无回答误差、计量误差做了讨论,提出了处理这些误差的一些方法。本章还对调查数据中离群值的检测和处理方法进行了讨论。需要注意的是,不论对于哪一种类型的非抽样误差,都有若干种处理方法,但任何一种方法都有它的使用条件和局限性。所以在应用中需要具体情况具体分析,简单照搬往往难以收到理想的效果。

习 题

1. 有一本几年前某地区居民住址的名录,上面有各条街道中居民住户的地址和户主姓名。现在想对该地区的居民进行一次入户抽样调查,这本名录抽样框有什么缺点?你打算怎样补救?

2. 欲对市场中的个体商贩进行抽样调查,有两个抽样框可供使用。一个是个体商贩在工商局注册的名录框,一个是市场摊位的地址框。这两个抽样框各有什么特点?你打算使用哪个抽样框?请说明你的理由。

3. 请就下面问题进行讨论:

(1) 调查中的无回答是怎样产生的?

(2) 你认为各种类型的无回答对估计会产生什么影响?请举几个例子说明你的观点。

(3) 对你所列举的无回答有没有比较好的预防措施?如果有,请就这些措施的操作性进行讨论。

(4) 如果出现了上述无回答,采用什么措施进行补救效果较好?

4. 假定在现场调查中,由于采用深入程度不同的调查方法,可以得到不同的回答率,这些回答率分别是 60%, 80%, 90% 和 95%。对于一个待估的百分比,各回答层真实均值如下表:

按回答率(%) 分层	真值(%)
60	40.7
80	43.5
90	44.8
95	45.4
5% 无回答层	59.0

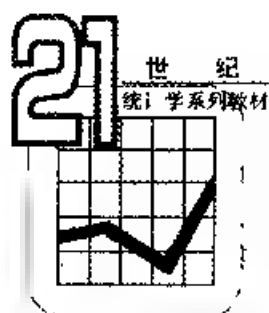
请回答下列问题:

(1) 采用最简单的调查手段只得到 60% 的回答率,在这种情况下,证明对整个总体百分比估计的均方误差的根是 $\sqrt{\frac{2414}{n} + 28.94}$, 其中 n 是回答者的数目。

(2) 证明当采用一个只能得到 60% 回答率的方法时,均方误差的根无法达到 5%,除非回答者稍微超过 100 人,或者回答率在 80% 以上。

(3) 如果均方误差的根规定为 2%,采用什么方法可以达到这个目的?需要多大的样本量?

5. 在上题(3)中,假定采用能得到 90% 回答率的调查方法,完成每份问卷的费用要上升到 \$5。如果从其余 10% 无回答中再得到一半(即总体 5%) 的回答,则每完成一份问卷的费用为 \$20。对于一个 2% 的均方误差的根,采用 90% 回答率的方法省钱还是采用 95% 回答率的方法省钱?



第 12 章

设计与方法 —— 美国 CPS 案例

美国人口现状调查(current population survey, CPS)被认为是全国性大规模居民住户抽样调查的典范。此项调查已有 60 多年的历史,但时至今日,每年仍有大量的论文讨论 CPS,这足以说明这项调查在人们心目中的地位,以及人们对事物完美的追求。作为教材的一个案例,这里简要介绍 CPS 的一些主要内容。第一节是 CPS 的概述,第二节介绍 CPS 的抽样设计,第三节介绍 CPS 的目标量估计,第四节介绍 CPS 的方差估计,第五节介绍 CPS 的非抽样误差及控制。

§ 12.1 概 述

一、背景

美国人口现状调查(CPS)是国际上最著名的抽样调查项目之一。之所以著名,一方面是因为其调查的内容多,调查对象分布的范围广,调查实施的难度大;另一方面在于该项目科学精巧的设计和系统有序的操作管理。此项目调查不仅为政府有关部门、社会科学家和各界人士提供了美国劳动力市场的全面信息,而且成为美国以及许多其他国家进行居民调查的参考模型。

CPS由美国联邦普查局(U.S. Census Bureau)和美国劳工统计局(U.S. Bureau of Labor Statistics)联合组织,它的核心数据是劳动力市场的信息,提供包括失业率、就业状况、行业收入等许多领域的数据。CPS还收集了大量人口数据,这有助于进一步了解按种族、年龄、性别等分类的全美国、各州以及州内不同地区、不同人口团体劳动力市场的状况。CPS由联邦普查局负责执行,使用的样本是经过科学方法挑选的5万多户居民。调查实施在每个月包含19号的那一周进行。调查的问题涉及受访者前一周(即包含12号的那一周)的活动。调查范围覆盖全美50个州和哥伦比亚地区。每位受访者连续4个月接受访问,然后退出样本,8个月后再进入样本,连续4个月接受访问,最后永久地退出样本。这种4-8-4的样本轮换设计保证了数据在月份之间、年度之间具有较高的相关性和可比性,可以反映样本的时序变化,又不断补充了新样本,避免了样本老化带来的诸多负作用。

CPS受访者必须是16周岁以上(含16周岁)的美国居民,因为16周岁以下人口的就业受到义务教育和儿童劳动法的限制,所以劳工统计局只出版16周岁以上(含16周岁)人口的劳动力资料。年龄没有上限,同等对待全日制学生与非学生。通常家庭中的一个成员代表家中所有成员接受调查,如果受访者对家庭其他成员就业情况不了解,调查员就必须与其他成员直接取得联系。

除了常规性的有关劳动力人口情况调查之外,通常CPS中还包括一些劳动力市场的分析家们感兴趣的问题,如兼职活动与收入、服役状况、受教育状况、暂时性的就业、工种更换、工作期以及其他一些内容。由于CPS样本量大、人口覆盖面广,因此许多赞助商利用该调查搜集一些其他数据,如对家庭大小的期望、吸烟状况、计算机使用状况以及选举投票情况等。这些调查内容以附加问题的方式插入到不同月份的调查中。

CPS的调查问卷是一份完全的电子版文件,数据采集方式有面访和电话调查两种。联邦普查局在马里兰州的黑格斯顿(Hagerstown Maryland)、印第安纳州的杰斐逊维尔(Jeffersonville Indiana)和亚利桑那州的图森(Tucson Arizona)建有3套数据收集系统中心,主要负责进行计算机辅助电话调查,面访调查则由访问员在便携式电脑上实施。

CPS始于1940年,这中间也经历了一些变化

二、CPS 历史沿革

美国国民失业状况如何,是美国政府和经济学家们十分关注的问题之一。在20世纪30年代经济大危机期间,测定失业率的问题变得更为突出。美国曾尝试用许多方法估计失业率,这中间还包含了相当成分的猜测。1937年,有关部门首次尝

试使用概率抽样的方法估计失业率。项目管理部门先是在地区范围,然后在全国范围研究并发展了测定失业率的技巧。这些研究为1940年的失业样本调查奠定了基础,从1940年起,失业样本调查成为项目管理部门每月一次的工作,可以认为这是CPS的开始。下面以10年为一个阶段,介绍CPS发展过程。

(一)1940—1950年大事记

1942年8月。联邦普查局接手失业样本调查的工作。

1943年10月。联邦普查局对样本进行了全面改动,改动后的住户样本来自68个初级抽样单元,覆盖了125个县和市。到1945年,约有25 000个住户进入样本。

1945年7月。对CPS的问卷进行了修改,增加了4个就业状况的基本问题,对一些调查项目重新定义。

1947年8月。对样本抽选方法做了修改,实现了在地区样本中每个单位入选样本的概率相等,从而简化了制表与估计过程。

1949年7月。进一步扩大了样本涵盖范围,包括饭店、汽车旅馆、拖车移动房等,因为这些住所居民的特点与其他人口不太一样。这些改变提高了数据质量。

(二)1950—1960年大事记

1953年1月。对目标变量采用比率估计的方法,比率估计的辅助变量为1950年的人口普查数据。新的估计方法的使用进一步提高了估计效率。

1953年9月。高速电子设备引入数据处理和制表过程,这不仅大大提高了估计的速度,同时带来估计方面的改进。电子设备的采用还扩大了抽样变异系数计算的范围。1959年,CPS采用现代计算机,从此联邦普查局就不断根据计算机的发展对调查的计算机环境进行同步更新。

1954年2月。初级抽样单元的个数由68个扩大到达230个,而样本住户25 000个的数量保持不变。同时调查的估计过程也有所改进,复合估计利用月份之间调查样本的重合优势,在没有增加样本量的情况下,提高了绝大部分主要数据的可靠性。

1956年5月。初级抽样单元的个数由230个扩张到330个,样本覆盖了638个县和市,样本量也由原先的25 000个住户增加到40 000个住户。据测算,样本量的扩大使主要数据的可靠性提高了大约20%,并获取了更详尽的数据。

1957年1月。对就业状态的定义进行了重新的修正。按照新的定义,原先一些属于就业的人群被归入了失业人群。

1957年6月。调查加入了季节性调整,在随后年度内联邦普查局和劳工统计局进行的调查中出现了方法上的极大发展。

1959年7月。CPS的任务在不同机构之间进行划分。计划、分析及出版CPS劳

动力数据的任务由劳工统计局承担。数据采集、计算机数据加工、样本维护以及相关方法的研究任务由联邦普查局承担

(三)1960—1970 年大事记

1960年1月 阿拉斯加和夏威夷进入CPS样本,这使得初级抽样单元由原先的330个增加到333个。这两个州的加入增加了新数据与旧数据进行比较的难度

1961年10月 采用了计算机输入用胶片光学扫描装置(FOSDIC),CPS的调查问卷变成了1960年人口普查采用的FOSDIC格式。利用这个系统可以对调查结果直接扫描,并将信息传输到计算机中储存。这个系统可以使问卷的编排更加灵活、包容更多的问题。这个系统一直延用到1993年12月。

1963年3月。对比率估计中使用的样本与人口数据进行了调整,从而反映了1960年以来人口数量与人口分布的变化。根据人口分布变化,初级抽样单元逐步增加到357个,给人口增长迅速的地区以更充分的覆盖率。虽然调整后的总样本量不变,但使大多数数据的可靠性提高了5%。

1967年1月,将原来的357个初级样本进一步扩大到449个。样本量由原来的40 000个住户增加到60 000个。这次样本量的扩大使主要数据的可靠性提高了约20%,同时对就业与失业的概念进行了重新修订。修订包括对劳动力年龄的改动,增加了有关工作时数、失业时间及个体经营状况等问题。失业定义的修订使时序变化的估计产生了一些波动。

(四)1970—1980 年大事记

1973年3月。为了优化样本,初级抽样单元由449个进一步提高到461个,但样本量却由原来的60 000个减少到58 000个。这种改变在于,最终住户群中包含的住户个数从6个相近(但不相邻)变成4个相邻的住户。

1975年9月。采用了州追加样本。追加样本包含165个新的初级抽样单元约14 000个住户,用以补充26个州和哥伦比亚地区的国家样本。追加样本的目的是满足各州年平均失业人数估计的可靠性要求。1976年8月,重新改进了估计过程,并重新修订了可靠性要求,这样使三个州的追加样本退出,最后,追加样本包括155个初级抽样单元中的11 000个住户。

1979年1月。采用了新的两阶段比率估计程序。对家庭成员间关系、种族等方面的数据采集方式也进行了调整,例如,家庭成员的种族由受访者而不是访问员决定

(五)1980—1990 年大事记

1981年5月。1980—1981年间,对样本量又陆续做过一些调整。到1981年5月,初级抽样单元的总数达到629个,样本总量为72 000个住户。

1982年11月。增加了有关工会的调查问题。

1984年9月。开始收集女性服役情况的数据。

1984年10月。针对16岁—24岁的人口增加了入学情况的问题。

1985年6月。马里兰州的黑格斯顿开通了计算机辅助电话调查系统(computer assistant telephone interview, CATI)。在接下来的几年中对该系统进行了测试。

1987年4月。CPS首次月度估计中使用通过CATI系统中心得到的数据(亚利桑那州的CATI系统中心建于1992年5月,印第安纳州的CATI系统中心建于1994年9月)。

(六)1990年以后的大事记

1990年6月。一系列首次进行的新的劳动力问卷测试在黑格斯顿CATI系统中心展开。1990与1991年进行的这些测试使用了随机拨号技术。

1992年7月。CATI与CAPI(计算机辅助面访调查,即访问员携带笔记本电脑实施调查)综合试验开始进行。

1994年1月。CPS开始采用一套全新的专为计算机辅助访问设计的问卷。

1994年12月。对与受访者一起居住但非正式家庭成员的这种关系进行了新的分类,涉及未结婚的同居伙伴、室友及住客等。

1996年1月。对CPS的抽样设计进行了一些改动,原来的可靠性要求放松了。修改后的抽样方案中要求有754个初级抽样单元,而样本总量减少到50 000个住户。

综上所述,自1940年CPS开始正式实施,半个多世纪以来其改进、充实、完善的工作一直没有间断,不断追求、不断发展的线索清晰可见。概括起来,可以归纳为以下几个方面。

1. 调查内容不断完善。CPS对主要目标变量的定义进行过若干次修改,使得新定义能及时反映不断变化的经济环境,并保证更好的可操作性。调查的内容也随着时代的要求不断丰富,形成比较科学的指标体系。

2. 抽样设计的效率不断提高。CPS抽样设计演变的总趋势是,在多阶段抽样中,初级抽样单元的数目在不断增加,而最终样本量(住户)的数目却保持不变,甚至减少。根据抽样原理,在多阶段抽样中,第一阶段的样本方差在总抽样误差中占有重要地位。CPS抽样设计的不断改进,印证了抽样理论的指导作用。一阶样本的合理分布,以对初级抽样单元的分层为依据。所以在多阶段抽样中,对初级抽样单元的精细分层是非常重要的。

3. 估计方法精益求精。CPS在估计方法上也是精益求精。如该调查所采用的二阶段比率估计、复合估计、方差估计等都是众多专家、学者集体智慧的结晶。美国

政府对这方面的研究给予了很大的财力资助,许多论文也是以 CPS 中的问题作为讨论的背景

4. 采用高科技的设备与技术 CPS 总是以最快的速度把最新的高科技装备和技术应用于调查活动的实践中,如对光学扫描装置 FDSIIC 的使用,CAPI 系统中心的建立以及 CAPI 和 CAPI 的结合运用 在如此大规模的全国性调查中率先引入先进设备与技术,是 CPS 的一个亮点

5. 重视调查结果的评估 在调查方案经过较大调整、修订后,都有很详细的质量效果评估,以使对改进所带来的收益有一个清楚的认识,这项工作对于总结经验、积累素材也是非常重要的

§ 12.2 CPS 抽样设计

一、概述

50 多年以来,CPS 一直是美国劳动力与人口特征方面最新信息的主要来源 因为 CPS 的重要性与高层次性,对它的可靠性评估定期进行 伴随美国 10 年一次的人口普查,CPS 抽样设计也是 10 年修订一次,修订通常在两次人口普查中间,新的抽样设计尽可能多地利用人口普查提供的信息,同时兼顾到两次普查之间人口状况的变化 最近一次的抽样设计于 1995 年 7 月完成。由于经费下调,1996 年 1 月 CPS 抽样设计又经过一次调整,但主要是对某些州样本量的调整,抽样设计的思想与方法没有改变。本节所介绍的内容,取自 1995 年 7 月的抽样设计。

CPS 抽样设计具有以下几个主要特征:

1. CPS 样本是随机样本。
- 2 调查的核心内容是 16 周岁及 16 周岁以上家庭人口的劳动力特征。
3. 抽样时以州为总体,因而设计也是以州为总体的设计。事实上,各州的抽样方案都是统一的,区别在于各州对核心变量估计精度的要求不同,因而样本量不同 劳工统计局和联邦普查局负责总的计划与协调,并根据各州调查结果对全国数据进行推估。
4. 样本量由变异系数 CV 及可靠性要求所决定。变异系数是衡量抽样误差的一个相对数,它等于估计量标准差除以变量的期望值 就全国而言,通常假定失业率的期望值为 6%,变异系数要求为 1.8%,在显著性水平 $\alpha = 0.1$ 条件下,对全国失业率估计的误差范围在 $\pm 0.2\%$ 之间。

5. 在失业率为 6% 的自定义下,各州对变异系数的要求在 8%—9% 之间,这

样就能保证进行全国估计的变异系数控制在 1.8% 之内

CPS 抽样的主体部分是采用二阶段抽样。就全国范围而言,第一阶段采用分层 PPS 抽样,抽出 754 个初级抽样单元 (PSU),第二阶段采用整群系统抽样抽出最终包括 56 000 个住户的样本。有时,当实际产生的最终样本单位过大,就需要第二阶段的抽样。抽样设计保证在一州内绝大多数住户最终被选入样本的概率是相同的,但是由于设计是以州为单位的,所以不同州的住户最终被抽中的概率是有区别的。当然,如果只考虑国家水平的数据,更有效的设计方案或许应该使全国所有住户被抽中的概率相同,但那样就无法保证州水平与国家水平数据的可靠性同时得到满足。因此,目前的这种设计兼顾了国家和州两级的需要。

二、第一阶段的抽样

第一阶段的抽样涉及三个方面的工作。这些工作是:初级抽样单元 (PSU) 的界定;将初级抽样单元 PSU 分层;PSU 的抽选。

(一) PSU 的界定

PSU 是不跨州界的。组成 PSU 的基本行政区划是县,但也不是绝对的。初级抽样单元 PSU 或者是一个县,或者是相邻的两个或多个县,在城市,PSU 按照城市统计区域 (metropolitan statistical area, MSA) 界定。对每个 PSU 的要求是,面积不超过 3 000 平方英里 (相当于 7 770 平方千米),人口在 7 500 人以上。如果面积与人口数发生冲突,例如在人口稀少的地区,3 000 平方英里的范围内人口低于 7 500 人,则在 PSU 界定时面积具有优先权。这主要是保证每个 PSU 的地理范围不能过大,以保证访问员的实际操作。美国的 PSU 规则产生于 20 世纪 40 年代末期,后来对规则不断调整。上述所言为 1990 年 PSU 规则的主要内容。根据上述要求,目前美国的 3 141 个行政县共划分为 2 007 个初级抽样单元 PSU。

(二) 对 PSU 的分层

对 PSU 进行分层的主要标准有两个:一个是在同一层内,各 PSU 具有很大的相同特征;另一个是各层的规模接近,即每一层中的人口数接近。有些 PSU (如城市),人口密度大,则这些 PSU 被归入必选的初级单元。这样就把 2 007 个 PSU 划分为两类:

第一类:具有自代表性质的 PSU (self-representing), 共 432 个。这 432 个 PSU 是必选的初级单元。

第二类:非自代表性质的 PSU (non self-representing), 共 1 575 个,这中间的样本单元是通过随机抽选产生的。

将 1 575 个非自代表性质的 PSU 按地理位置 (州内)、人口统计学特征和人数

规模分为 360 个层,平均每层中约有 4 - 5 个 PSU。

(三) PSU 的抽选

每个具有自代表性质的 PSU 自然进入样本。这样,在第一阶段的初级抽样单元中,共有 432 个自代表的 PSU 在其他 360 个层中,采用与人口规模成比例的概率抽样,从每个层中抽取一个 PSU。于是,第一阶段抽样中共抽取出 792 个(1996 年又减少到 754 个)初级抽样单元。

三、第二阶段的抽样

CPS 基本上是采用二阶段的抽样,故第二阶段抽样实际上是抽取最终抽样单元(USU)抽选时采用整群抽样方法,每个 USU 由 4 个住户住址所组成,大多数情况下,这些住户(即住房的地址)都是独立的家庭单位。然而,随着时间的变迁,一些房舍可能被拆毁或者被转为其他非居住用;有的住户地址可能由几个家庭所分用。这些住户地址仍然是抽样单位,但这些情况会使一个群的大小发生一些微小的变化。通常,4 个相邻的住户地址组成一个群,有时这些住户地址也会比较分散,但与其他住户地址相比,构成一群的 4 个住户地址应当是最为邻近的。这样做的好处是便于实施调查,节省调查费用;其弱点是,由于相邻的住户可能具有较多的相似性,因而会增大抽样误差。

在美国,将生活区域分为两大类:一类为居住单位或住户。一个住户是指有一套房间或一个单独房间作为一个独立的生活区,他们与其他生活区通过如公寓楼的大厅和走廊发生关系。在一个住户中居住的或者是一个人,或者是一个家庭(这是绝大多数情况),或者是两个或两个以上没有家庭关系的人。在 1990 年的人口普查中,有大约 98% 的人口居住在这样的住户中。另一类是集体户。集体户是指居住者共同享用公共设施或得到统一的照顾,例如学校宿舍、养老院、福利社等。在 1990 年的人口普查中,有大约 2% 的人口居住在集体户中。

二阶段抽样时,使用的抽样框主要有三个:(1)集体户抽样框;(2)住户抽样框;(3)区域抽样框。下面简要介绍三个抽样框的构造。

(一) 集体户抽样框

在抽样设计中,每个 USU 只包括 4 个住户(家庭)单位,所以首先要将集体户人口转化为住户抽样单位。转化方式为,用集体户总体人口除以 2.63(1990 年人口普查时每个住户的平均人口为 2.63 人),然后将转化的 4 个住户单位组成一群。

(二) 住户抽样框

住户抽样框由有完整地址的住户单位所构成,典型的完整地址有街道名称和门牌号,如“榆树街 1599 号”。大多数情况下,每个住户地址都是一个独立的家庭单

位,4个邻近的住户地址组成一群。抽样框由群排列而成,采用系统抽样方式抽取群。

(三) 区域抽样框

住户抽样框难以包括所有的住户单位,如有些住户没有完整的地址,或只有邮寄地址而没有确切的登门地址,如“PO123信箱”。随着时间变迁,也会有一些新住户出现而没能反映在住户抽样框中。所以区域抽样框是住户抽样框的补充,它包括那些地址不确切的住户,也包括从建筑许可部门所获得的有关新建筑的信息。

最终抽样单元 USU 的抽取也是由各州独立进行,抽选时是以 $1/k$ 的抽样概率从每个初级抽样单元 PSU 中抽取系统样本。这里 k 是 PSU 内的抽样间隔,由于各州的抽样比不同,因此各州 PSU 中的 k 值是不同的。但对于同一个州而言,不同家庭最终入选样本的概率是相同的。由于 CPS 的抽样设计是 10 年修订一次,在这 10 年期间,为了保证样本轮换,抽选时将 10 年间准备轮换的样本一并抽出备用。

最后需要补充的是,有时最终抽样单元的大小与设计要求有所偏离,这些偏离会影响到调查员工作的顺利完成。所以,如果当最终抽样单元有 15 个以上的住户单位时,就需要采用第三阶段抽样。在一系列的工作实施后,工作人员会摸清这种情况,并将原先的抽样单元划分为若干个更小的最终抽样单元,并在此基础上进行第二个阶段抽样。由于三阶段抽样改变了住户单位被选中的概率,所以,如果出现这种情况,在进行估计时需要使用加权因子对抽样概率进行调整。

四、样本轮换

CPS 的样本轮换采用的是 4-8-4 模式,即一个住户单位在连续的 4 个月内接受调查,在接下来的 8 个月中退出样本,然后再接受连续 4 个月的调查,最终退出样本。轮换方案的设计使得具有相同特征的住户单位替换退出的住户单位。

CPS 的样本轮换具有以下主要特征:

1. 在任何一个月内,都有 $1/8$ 的住户单位第一次接受调查, $1/8$ 的住户单位第二次接受调查,如此下去。
2. 每个月都有新的样本组代替从样本中永久退出的老样本组。
3. 每个月都有一个样本组在 8 个月的闲置后重新接受调查。重新接受调查的样本组代替了刚刚退出、进入闲置期的样本组。
4. 轮换设计保证了每个样本单元在 2 个年份的 4 个相同月份中接受调查。
5. 在连续的 2 个月内,有 $3/4$ 的样本是相同的;在连续的 2 年中,有 $1/2$ 的样本是相同的。

前面提到,CPS 的抽样设计大体上是 10 年修订一次。新的抽样方案涉及对初

级抽样单元 PSU 的重新界定、PSU 的样本数目的改变以及对 PSU 的重新抽取。这样,就要在新入选样本的 PSU 地区雇用新的调查员,而且重新设计的抽样方案也往往会对调查程序作一些修订。于是,前后两个方案的样本衔接就是一个需要注意的问题。新方案的样本是逐步引入 CPS 实施过程中的,以保证调查过程的连续和调查数据的衔接。事实上,新的抽样方案的实施从 1994 年 4 月就已经开始,经过一年多的时间,到 1995 年 7 月彻底完成。

§ 12.3 CPS 目标量估计

一、概述

CPS 是以住户为单位,对全美国进行的一阶段抽样调查。它所估计的主要目标量是以劳动力资源为特征的一系列统计指标,包括人口总数、性别、年龄、种族的分布等。为了从调查数据中得到各州和全国的估计数据,就需要对样本中的每个被调查单位进行加权。从技术角度看,如何确定权数是目标量估计中的核心问题。

CPS 目标量估计程序中的权数确定,大体需要经过以下步骤:

1. 确定 CPS 样本的基础权数和特殊权数;
2. 根据无回答情况对样本权数进行调整;
3. 为减少 PSU 样本方差进行第一阶段比例调整;
4. 为进一步提高估计效率进行第二阶段比例调整;
5. 结合以前月份的调查数据进行复合估计,以进一步减小方差

二、基础权数和特殊权数

CPS 采用的是概率抽样,概率抽样可以得到目标量的无偏估计。为了得到目标量的无偏估计,需要用每个样本单元的调查值乘以该单元入选样本概率的倒数,然后汇总这些结果即可。

所以,基础权数即各单元入选样本概率的倒数。例如,某样本单元入选的概率为 1%,则该单元的权数为 1 000,意味着该样本单元的情况代表了 1 000 个单元的情况。由于抽样以各州为总体,抽样设计采用与规模大小成比例的自加权设计,所以在各州内,各样本单元的基础权数是相同的。当然由于各个州的抽样比不同,基础权数在各个州是不同的。

特殊权数是对发生在抽样最终单元 USU 时出现特殊情况而对权数进行的调

整。上一节已经谈到,对每个USU而言,期望所包含的住户为4个。但调查实施时发现,该USU包含的住户不是4个,比如说是8个。如果在这个USU中仍然只调查4户,对每户就要予以一个特殊权数2。于是该USU中每个被调查户的权数就是基础权数乘以特殊权数。当然,特殊权数的出现会给方差估计带来一些负面影响。为此规定,特殊权数被限制在4以内。

三、无回答调整

调查中会出现无回答的情况。无回答有两种类型:单元无回答和项目无回答。单元无回答指被访户没有接受调查,如被访户家中无人或拒访等。根据历史资料,在CPS调查中单元无回答率每月约为4%—5%,目前这一比率还有上升趋势。项目无回答指被调查户不能或拒绝提供某一个问题的信息。在数据处理过程中,对项目无回答有进行处理的专门程序,一般是采用插补的方法为缺失值模拟一个替补值。所以,这里所讨论的权数调整主要是针对单元无回答而言。

为了进行权数调整,需要构造调整层,使得在同一层内的回答单元和无回答单元的背景尽量相似。调整层是在初级抽样单元PSU的基础上进行的。首先将每个州的PSU分为两类,大城市类和非大城市类,在每一类中又分为两个调整层。大城市类可分为中心城市与非中心城市,非大城市类可分为城镇与农村。此外还有一些另类地区单独分层,这样全国共分为254个调整层。

对于每个调整层,分别用表格列出受访户与无回答户的权数,这个权数是基础权数乘以特殊权数。然后计算无回答调整系数。调整系数的计算公式为:

$$F_{ij} = \frac{Z_{ij} + N_{ij}}{Z_{ij}} \quad (12.1)$$

式中, Z_{ij} 为第*i*类第*j*层接受调查户的权数总和; N_{ij} 为第*i*类第*j*层无回答户的权数总和。

当调整系数大于2时,即接受调查户的权数不及总权数的50%时,需要将该类两个层的加权总数合并计算,目的是希望在某一层出现较高的无回答率时,调整系数仍具有较好的稳定性和代表性。历史数据表明,这种合并的情况是不多的。

至此,每个受访户的权数为:

基础权数 × 特殊权数 × 无回答调整系数

四、第一阶段比例调整

一些典型的人口特征与劳动力数据密切相关,这些情况包括年龄、种族、性别。

美国最重要的两个种族(除白人外)是黑人和西班牙裔,他们的情况受到格外的关注。调查中每个月 CPS 样本的人口特征分布与总人口的真实分布有所不同,IPS 用加权处理,使得在这些特征上样本的人口分布尽可能接近已知的总人口分布,实现的途径是采用比例调整。CPS 估计过程中有两个比例调整:第一阶段比例调整 and 第二阶段比例调整。第一阶段比例调整主要是对样本中黑人分布进行的调整,通过调整使得每个州内 PSU 中的黑人与非黑人比例接近该 PSU 所代表范围内黑人与非黑人的比例。在 CPS 抽样设计中,PSU 被分为两类,一类为肯定入选样本,这些 PSU 具有自我代表(self-representing)性质;另一类是通过抽选进入样本,这些 PSU 不具有自我代表(non-self-representing)性质(见第二节 CPS 抽样设计)。因此,第一阶段的比例调整上是针对不具有自我代表性质的 PSU 而言。

第一阶段比例调整因子采用下面公式计算:

$$FS_{sj} = \frac{\sum_{i=1}^n C_{sij}}{\sum_{k=1}^m \left[\frac{1}{\pi_{sk}} \right] C_{skj}} \quad (12.2)$$

式中, FS_{sj} 为 s 州中第 j 个种族的第一阶段调整因子(j = 黑人, 非黑人); C_{sij} 为 s 州中第 j 个种族第 i 个非自我代表 PSU 16 岁以上的人口总数; C_{skj} 为 s 州中第 j 个种族第 k 个非自我代表 PSU 16 岁以上的人口总数; π_{sk} 为 s 州中第 k 个非自我代表样本 PSU 入样的概率; n 为 s 州中非自我代表 PSU 的总数(包括入样的和非入样的); m 为 s 州中非自我代表 PSU 的样本个数。

但如果一个州内的黑人或非黑人调整因子满足下列条件之一,即将两个因子合并:(1) 因子大于 1.3;(2) 因子小于 $0.76923 \left(\frac{1}{1.3} \right)$;(3) 该州内少于 4 个非自我代表 PSU 样本;(4) 该州内某一种族少于 10 个受访者。

第一阶段比例调整后,每个受访者的权数为:

基础权数 \times 特殊权数 \times 无回答调整系数 \times 第一阶段比例调整因子

五、第二阶段比例调整

第二阶段比例调整的程序要复杂一些。它的基本思想是采用迭代的方法,将样本中一些有关人口特征的重要变量的权数调整到与总体数量尽可能一致。每个月 CPS 的样本都是由 8 个轮换组构成的,调整在每个轮换组中进行。进行调整时有三套控制变量,它们是:(1) 各州 16 岁以上公民的总数;(2) 西班牙裔/性别分组(共 14 组)和非西班牙裔/年龄分组(共 5 组);(3) 白种人/性别/年龄分组(共 66 组),

黑种人/性别/年龄分组(共42组),其他种族/性别/年龄分组(共10组)。

由种族、性别和年龄变量组合成的分组的个数不同,体现了设计人员对不同方面内容关注程度的不同。

在权数调整中,如果仅以一套控制变量为目标,势必引起与其他控制变量的偏离,因此需要采用迭代方法使调整权数同时适应所有控制变量。一般而言,经过6次迭代,就可以达到权数调整的目的。研究人员认为,第二阶段比例调整不仅可以减小CPS估计误差,而且当迭代收敛时,估计量可以使下面的统计量最小化:

$$\sum_i W_{2i} \ln \frac{W_{2i}}{W_{1i}} \quad (12.3)$$

式中, W_{2i} 为第 i 个样本单元的最终权数; W_{1i} 为第 i 个样本单元的第一阶段比例调整后的权数。

进行第二阶段比例调整需要大量辅助信息,这些数据来自于人口普查和其他有关渠道。

由于几个控制变量的多种组合,形成了众多的调整组。如果某些调整组中没有受访者,或受访者过多,都可能增大样本估计的方差,因此对这些情况需要进行识别。于是,在迭代之前,首先需要计算初始调整因子,计算公式为:

$$\text{初始因子 } F_{jk} = \frac{C_j}{E_{jk}} \quad (12.4)$$

式中, C_j 为第 j 个调整组中的控制总量除以8(因为CPS样本中有8个轮换组,调整是在每个轮换组中分别进行); E_{jk} 为第一阶段比例调整后第 j 个调整组中的第 k 个轮换组对控制变量的估计。

这些初始因子的作用是决定调整组是否需要合并,如果符合下列条件之一,则该组与邻近的调整组合并:(1) 调整组中没有受访者;(2) 初始因子小子或等于0.6;(3) 初始因子大于或等于2。

每一次的迭代可以分解为三步:计算组内的调整因子,用调整因子进行组内的估计,用控制总量除以该估计值。这些步骤重复6遍(即6次迭代)得到第二阶段比例调整因子。在这个基础上,求出每个受访者的最终权数。

$$\text{最终权数} = \text{基础权数} \times \text{特殊权数} \times \text{无回答调整系数} \times \text{第一阶段比例调整因子} \times \text{第二阶段比例调整因子}$$

六、复合估计

一旦获得样本单元的最终权数,采用霍维茨-汤普森(Horvitz-Thompson)估计量,就可以得到CPS估计值。但在CPS中,对有关劳动力资源的大多数目标量,

采用的是复合估计(composite estimate)方法。复合估计是把几个估计值加权平均。CPS的复合估计包括了以下内容,一个是上面所提到的采用霍维茨-汤普森估计量所得到的CPS当前调查月份的估计,另一个是上月份的复合估计以及对两个月份之间变化量的估计。变化量估计的数据来源于两个月份中75%的相同样本。在1985年之前,复合估计只采用当月的霍维茨-汤普森估计量和上月份的复合估计量,并对这两个估计量赋予相同的权数。1985年后估计方法有所改变,对上述两个估计量分别赋予不同的权数,并补充了两个月份之间变化量估计这个因子。以劳动力水平为例,对劳动力水平 Y 进行复合估计的公式为:

$$Y_t = (1-k)\hat{Y}_t + k(Y_{t-1} + \Delta_t) + A\hat{B}_t \quad (12.5)$$

这里

$$\hat{Y}_t = \sum_{i=1}^8 X_{t,i}, \Delta_t = \frac{4}{3} \sum_{i=1}^8 (X_{t,i} - X_{t-1,i}), \hat{B}_t = \sum_{i \in S} X_{t,i} - \frac{1}{3} \sum_{i \in S} X_{t,i}$$

$i = 1, 2, \dots, 8$ (8个轮换组)

$X_{t,i}$ 为第 t 个月份第 i 个轮换样本第二阶段比例调整之后目标量的权数之和;

$S = 2, 3, 4, 6, 7, 8$

式中, \hat{Y}_t 为目标量当前月份的霍维茨-汤普森估计; Δ_t 为月份 $t-1$ 到月份 t 轮换组变化量的估计; \hat{B}_t 为样本中新进入的轮换组(第1,5组)和原有的组(第2,3,4,6,7,8组)之间净差异的估计值。在估计公式中之所以加上此项,是因为历史数据表明,样本中的新进入组与原有组相比,在有关劳动力资源和失业率方面往往表现出趋高的倾向,加上这一项可以在某种程度上起平衡作用。如果不存在上述倾向, \hat{B}_t 的期望值则为零。

有关CPS估计的研究还表明,当取常数 $k=0.4, A=0.2$ 时,对有关劳动力特征的变量来说,估计量的方差可以下降到最理想的程度(参见Kostanich, D. and Bettin, P.(1986), "Choosing a Composite Estimator for CPS", presented for presentation at the International Symposium on Panel Surveys, Washington, DC)。

§ 12.4 CPS的方差估计

一、概述

CPS中的方差估计主要用于两个目的:一是对估计值的方差进行估计,以用于

各种统计分析; ②是对每一阶段抽样效果和估计的精确度进行评估,以评价和改进抽样设计

本节中所讨论的问题主要有:方差估计的再抽样方法;1990年抽样设计的方差估计方法;州和地区水平的方差估计;广义方差;用估计方差评估抽样设计。

二、方差估计的再抽样方法

用再抽样方法进行方差估计是复杂样本方差估计经常采用的方法。这种再抽样是在每个月的总样本中,采用与总样本相同的抽样原则和估计程序,抽取一些次级随机样本,根据这些次级随机样本的估计值计算方差。可以把这些次级样本称为重复样本。增加重复样本的数量会提高方差估计值的准确性,但也会因此而增大费用。所以,重复样本的数量由调查成本和对方差估计质量要求此消彼长的关系决定。

1970年以前,CPS的方差估计使用40个重复样本,受计算机功能的限制,仅计算了14个特征组的方差。1970年的设计采纳了Keyfitz方差计算方法,这些方差估计运用泰勒级数法,消掉含有比第一项更高阶导数的项。到1980年,计算机内存的改进,使得采用复合估计对所有阶段进行加权的再抽样,计算多个目标量的方差估计值成为可能。

1980年以后,开始使用均衡半样本方法进行方差估计,样本被分割成48个重复样本。重复样本保持了抽样设计的所有特征,如分层方法、PSU内的样本抽取方法等。但由于成本费用和计算机的限制,这种方差估计的方法仅仅使用了13个月(1987年1月到1988年1月)。而且由于分解以后有些PSU的规模过小,地方一级水平上的方差估计被认为不够可靠。

三、1990年抽样设计的方差估计方法

1990年设计方案中的方差估计采用了逐次差分再抽样法。该方法的理论基础由Wolter(1984)提出,又由Fay和Train(1995)进一步发展。该方法是将被选中的最终样本单元USU(一般包含4个住户单元)按相邻的顺序配对,如(USU1, USU2), (USU2, USU3), (USU3, USU4)等。这种方法在方差估计中更好地反映了系统抽样的特点,重复样本的数量也由48个增加到160个,以期提高估计的精度。

由方差理论知道,总方差由组内方差和组间方差两部分组成。在这里组内方差是初级抽样单元PSU内由于抽样所产生的误差,又称PSU组内方差。组间方差是

指在所有的非自我代表单元 PSU (简称 NSR PSU) 之间抽取 PSU 样本所产生的误差, 又称 PSU 组间方差。对于自我代表单元 PSU (简称 SR PSU), 因为其入样的概率为 100%, 不涉及该层次上的方差计算。所以逐次差分再抽样法只是针对总方差和 PSU 组内方差计算而言, PSU 组间方差则可以根据总方差和 PSU 组内方差值推算。

再抽样因子是根据一个 160×160 的 Hadamard 正交矩阵计算出来的。要估计总方差, 对于 SR 样本和 NSR 样本, 重复样本的组成方法是不同的。在州内, 由 NSR PSU 组成虚拟层, 虚拟层中的 PSU 被以随机的方式分配给重复样本的每个组。用再抽样因子 1.5 或 0.5 来调整 NSR 组的权数。这些因子由 Hadamard 矩阵的行来确定, 从而进一步解释了虚拟层中的最初层为何容量不同 (参见 Wolter (1985) "Introduction to Variance Estimation" New York, Springer Verlag)。大多数情况下, 虚拟层是由成对的 PSU 组成, 但在有些州, NSR PSU 个数为奇数, 就需要构成一个包含三个 PSU 的虚拟层, 在这种情况下, Hadamard 矩阵的两行被赋予该虚拟层, 使得三个 PSU 的再抽样因子为 0.5, 1.7 和 0.8, 或者 1.5, 0.3 和 1.2。一个虚拟层中的所有最终单元 USU 被赋予相同的行数。

对 SR 样本, Hadamard 矩阵的两行被赋予给每对 USU, 以构造出再抽样因子 f_r :

$$f_r = 1 + (2)^{\frac{3}{2}} a_{i+1,r} + (2)^{\frac{3}{2}} a_{i+2,r}, r = 1, 2, \dots, 160 \quad (12.6)$$

式中, $a_{i,r}$ 是在一个系统样本中, 第 i 个 USU 对应的 Hadamard 矩阵中的数字 (+1 或 -1), 根据此公式可以得出再抽样因子为 1.7, 1.0 和 0.3。

依照上节所述的估计程序, 求得每个重复样本目标变量的估计值 \hat{Y}_r , 和整个样本目标变量估计值 \hat{Y} , 则方差估计值为:

$$v(\hat{Y}) = \frac{4}{160} \sum_{r=1}^{160} (\hat{Y}_r - \hat{Y})^2 \quad (12.7)$$

SR 样本中的再抽样因子 1.7, 1.0 和 0.3, 是根据 (12.6) 式计算得到的, 并能得到 (12.7) 式中的数字 4, 从而保证了对于 SR 样本和 NSR 样本公式是一致的 (参见 Fay, R. and Traim G. (1995), "Aspects of Survey and Model Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties", Proceedings of the Section on Government Statistics, American Statistical Association pp. 154-159)。

以上方差估计的方法也可以用来估计 PSU 组内方差。对于 SR 样本, 采用与总方差估计相同的再抽样因子; 对于 NSR 样本, 则采用与 SR 样本相同方式组成 USU

配对。这样,对于 PSU 组内方差,所有的 USU 的再抽样因子均为 1.7,1.0 和 0.3

四、地区和州水平的方差估计

对一些地区,如大城市地区的方差估计可以通过全国样本中的 SR 样本部分的方差估计方法得到。据估计,由于方法的改进,这些地区的方差估计值比相同样本量用 1980 年的方法得到的估计值更可靠。然而,对于受 NSR 样本影响较大的地区,方差估计却是难以解决的问题。

州水平的方差估计也是需要注意的。有一些州,仅仅包含了少量的 NSR 性质的 PSU,将它们配对组成虚拟层,减少了重复样本的数量,方差估计的可靠性问题将变得更加严重。而在另外一些州,NSR 覆盖范围的人口比重超过全国的平均水平,在这些州由 PSU 得到的方差估计可能变得更为重要。

此外,为估计方差而构造虚拟层,带入了新的层间方差,而这部分在抽样设计中是没有考虑的,由此导致了对真实方差的偏高估计,对总体(全国)目标量的估计值进行方差估计时,这种新产生的层间方差影响相对较小,但在州一级水平上的估计,这种方差就显得较大。这是在估计州水平的方差时需要说明的。

对地区和州水平上的方差估计的研究仍在继续,并取得了一些进展。当这些研究成果可以操作时,方差估计将会得到进一步的改善。

五、广义方差(generalized variance)

除一些例外,在出版的报告和公开的数据中,抽样标准误差是用广义方差函数(简称 GVF)计算的。广义方差函数是一个简单的模型,是估计量期望值的方差函数表达式,模型中的参数采用上而所讨论的方法进行估计。这些模型提供了计算估计标准误差近似值相对容易的方法。

需要回答的一个问题是,为什么不是根据调查数据计算每个估计值的标准差?实际上可以这样做,但个别标准差的作用是有限的,数据的使用者不可能据此预测所有结果之间的关系,而这种关系是数据使用者所感兴趣的。更重要的是,方差估计值是基于样本数据得到的,也具有其自身的方差。某个月目标量估计值的方差估计通常不如目标量估计值本身更精确,这意味着同一指标方差估计值在不同的月份差别会很大,或者在同一个月内有具有相关的不同指标的方差估计值大相径庭。因此需要一些使得这些方差估计值更稳定的方法来提高其可靠性。经验表明,使用广义方差函数可能会得到更稳定的方差估计值。

广义方差函数用于产生人口总量 x 估计值的估计方差,函数形式为

$$\text{Var}(\hat{X}) = aX^2 + bX \quad (12.8)$$

式中, a 和 b 为用最小二乘法得到的估计参数。

该模型的原理是假定 x 的方差可以表示为简单随机样本的方差与设计效应 (deff) 的乘积。设计效应 $deff$ 是指某一复杂抽样设计相对于简单随机抽样设计的效果。定义 $P = \frac{X}{N}$ 为具有 X 特征的人口比例, N 为人口总数, $Q = 1 - P$, 由 n 个样本单元得到总量估计值 \hat{X} 的方差为:

$$\text{Var}(\hat{X}) = \frac{N^2 PQ(deff)}{n} \quad (12.9)$$

又可将上式写为:

$$\text{Var}(\hat{X}) = deff \frac{N}{n} \frac{X^2}{N} + deff \frac{N}{n} X \quad (12.10)$$

令

$$a = \frac{b}{N}, b = \frac{deff \cdot N}{n}$$

则

$$\text{Var}(\hat{X}) = aX^2 + bX$$

我们取 $a = \frac{b}{N}$, 这里 N 为控制总量, 当 $x = N$ 时, 方差为零。

参数 a 和 b 用相对方差模型进行估计。相对方差模型为:

$$VX^2 = a + \frac{b}{X} \quad (12.11)$$

相对方差 VX^2 是方差除以估计期望值平方。通过一组相关的估计值和这些估计值的相对方差拟合模型, 得到系数 a 和 b 的估计。估计值相对方差则是采用逐次差分再抽样方法计算得到的。

模型拟合采用迭代加权最小二乘法, 权数是相对方差平方的倒数。采用这个权数是为了防止具有较大的相对方差项对 a 和 b 的估计产生影响。通常, 至少使用一年的数据用于此模型的拟合, 同时, 每个组至少有 20 个项估计值及其相对方差。需要注意的是, 使用模型来估计估计值的相对方差可能会引入一些误差, 因为模型可能对一些合理的极值进行了很大的修正。计算广义方差是为了估计不同月份间的变化及月度水平, 参数 a 和 b 会定期更新, 以便反映总体总量的变化和由于样本缩减造成的 $\frac{N}{n}$ 之比的变化。

当 a 和 b 确定之后, 就可以构造一个表格, 在表格中给出具体的估计结果。然而许多报告都给出了有关参数的列表, 使用者可以直接利用列表计算广义方

差的估计值,表 12.1 就是从就业和薪酬的列表中(美国劳工部提供)截取的例子。

表 12.1 月度数据标准误差计算参数

特征组	a	b
失业:		
总体或白人	0.000 015 749	2 464.91
黑种人	0.000 191 460	2 621.89
西班牙裔	0.000 098 631	2 704.53

月度估计值 \hat{X} 的近似标准差 $S_{\hat{X}}$ 可以由上表中 a, b 及以下公式得到:

$$S_{\hat{X}} = \sqrt{a \hat{X}^2 + b \hat{X}} \quad (12.12)$$

假设某个月城市劳动力中的失业人数为 6 000 000 ($\hat{X} = 6\,000\,000$),则由表 12.1 知:

$$a = 0.000\,015\,749, b = 2\,464.91$$

$$S_{\hat{X}} = 119\,000$$

月度城市失业人口估计值 90% 的置信区间是 5 810 000 ~ 6 190 000 ($6\,000\,000 \pm 1.6 \times 119\,000$) 之间。

六、用估计方差评估抽样设计

自 CPS 调查开始以来,抽样设计和估计方法发生过很多变化,这是因为 CPS 总是力求最大限度地优化所能得到的资源,最大限度地运用最新的技术。在不同历史时期对可靠性要求的不同,也导致了这些变化。要进行有效的抽样设计,运用估计方差进行评估是必要的。

前面曾提到,CPS 的总方差包括 PSU 组内方差和 NSR PSU 组间方差。实际上,对于大部分指标来说,PSU 组内方差已经解释了总方差的 90%,但不同指标之间是有差异的,例如,对于国内劳动力和非劳动力指标,方差几乎全部来自于 PSU 内的住户单元抽样;对于农业就业总人口和农业白人就业人口,PSU 组内方差解释了总方差的 60% - 70%;而农业西班牙裔就业人口和非农业西班牙裔人口 PSU 组内方差占总方差的约 80% - 90%。这至少表明 CPS 对初级抽样单元 PSU 的界定、分类、分层和抽样是非常成功的。

相对方差可以用于对估计程序和估计步骤的评估。使用相对方差比使用方差本身更有意义,因为估计的不同阶段可能既影响估计水平,也影响方差本身。估计

程序包括:确定基础权数、特殊权数,无回答调整,第一和第二阶段比例调整,以及复合估计方法的采用。例如如果经过加权和调整,采用无偏估计方法(即不是采用复合估计方法)失业率相对方差(变异系数的平方)的估计值为 3.590×10^{-4} ,但如果仅仅采用基础权数和特殊权数进行估计,相对方差则是它的 1.06 倍,如果再加上无回答调整因子,相对方差则是它的 1.05 倍,由此可以看出在失业率指标的估计上各个加权调整程序所起的效果。

从整个估计过程看,对于全国指标的估计,第一阶段的比例调整对相对方差的估计几乎没有产生什么影响。第一阶段比例调整的目的是为了降低州水平上的估计值方差,但是否实现了这个目标,其效果尚待检验。

然而第二阶段的比例调整效果却是明显的,它的引入大大减小了总方差,特别是对那些像年龄、性别、种族这种比例较高的分组,效果尤其明显。例如,国内劳动力中白人、黑人或西班牙裔的人数或者非农业领域中的就业者人数,若没有第二阶段的比例调整,相对方差将会提高几倍。但对于一些规模很小的分组,如农业中就业和失业的情况,第二阶段的比例调整没有明显的效果。

估计程序的最后一步是复合估计。复合估计是利用两个月份中 75% 的复合样本的有关信息改进对月份变化的估计。例如,西班牙裔失业人数复合估计的估计方差为 3.659×10^9 ,它是经过各阶段加权以后估计值估计方差的 92%,也就是说,在该指标中,复合估计使估计方差下降了 8%。此外,也可以利用设计效应 $deff$ 作为评估的另一个指标。设计效应的计算公式是(12.9)式,式中的 P 和 Q 由 6 个月的数据综合而成。就失业人口而言,复合估计的设计效应为 $deff = 1.229$ 。这说明,在样本量相同情况下,CPS 抽样设计(包括样本抽选、加权程序和复合估计)的方差比简单随机抽样的无偏估计的方差高接近 23 个百分点。另一方面,如果不是采用复合估计,设计效应则为 1.314,这也说明,在 CPS 设计中复合估计比通常的估计有更好的效果。

§ 12.5 非抽样误差及控制

一、概述

非抽样误差可能产生在调查的各个阶段,而且不易辨认。显然,非抽样误差的存在将影响调查结果中偏差的产生和方差的增大,但这些影响难以被测量。因此最适当的策略莫过于了解非抽样误差产生的原因,并在调查中采取有效的措施防范。

1994 年以前,研究人员对 CPS 中的非抽样误差曾做过很好的研究,但 1994 年

1月以后,对CPS中的非抽样误差没有像以前那样进行过精密的测定。由于CPS使用了计算机作为采集数据的工具,因而以前的研究结果对目前的情况在多大程度上适用也不清楚。但是可以肯定,有几类主要的非抽样误差一定存在于CPS之中。

一类是由于抽样框或其他信息资源有误所引起的误差,另一类是由于调查中的无回答所带来的误差,还有一类误差产生于被调查者不真实的回答。本节将对这几种类型的误差来源和控制措施做简要的介绍和讨论。

二、抽样框误差及控制

完善的抽样框的标志是,目标总体与被抽样总体中的单元一一对应,抽样框中的目标单元既没有重复,也没有遗漏。但调查实践中很难找到这样好的抽样框,在大规模调查中更是如此。在CPS中,两类问题比较明显:一个是遗漏,即抽样框中遗漏掉目标总体单元;另一个是不适当涵盖,即抽样框中包含了不应接受调查的单元。从CPS的历史看,遗漏情况更为严重,其后果是造成总量估计偏低。

遗漏产生的主要来源有:(1)地址不清。指调查地址不完整或有些单元无法确定其地址。(2)新建筑产生。特别是农村一些地区未经建筑许可部门批准而兴建的建筑。这些建筑没有被登记,因而不会出现在抽样框中。(3)户内单元遗漏。把本应属于某住户单元的成员错认为不属于该住户单元的成员。此外,还包括对无家可归者的遗漏。据测算,1997年1月CPS对16岁以上人口总体的遗漏率(包括各种类型的遗漏)大约为8%。一般来说,对黑人的遗漏率最高,约为17%,其中又以25岁~34岁年龄组的男性黑人遗漏率最高,对女性的遗漏率则要低一些。

抽样框误差的控制措施主要包括以下几个方面。

(一) 样本检验

样本检验包括样本测试和结果检验两个方面。

样本测试是在样本抽取之前对各种抽样程序进行测试,以保证这些程序可以单独或联合运行,用一些由极端值构成的小规模数据检验整个系统在特殊情况下的表现。例如,用1990年普查中的异常情况检验同年CPS抽样过程。这些情况包括航海器中的船员、特殊的集体户作以及印第安保留地中的情况等。设计人员编写了一些程序用以验证抽样过程是否能够正确、恰当地解决这些问题。

结果检验主要是对抽样结果的检验,验证系统是否是在真实数据的基础上运行。检验的例子有:初级抽样单元PSU入样概率之和是否为1;查验文件中的空字段及超出定义范围的数据;检验文件的一致性,例如选入样本的PSU与输出结果的内容是否相一致;检验信息的集中化程度,如所有各州的抽样比应纳入同一个参数文件中,等等。作为一项整体一致性的验证手段,抽样过程中某些阶段的结果被

用来与以前 CPS 调查设计的结果进行比较

(二) 名单审核

对每个月使用的名录进行审核,以把抽样框误差控制在尽可能小的程度。目前,名录审核工作的速度尚不足以保证工作过程中发现的错误都能及时得到纠正,但审核的自动化程度已使审核进度得到了巨大的提高,从而为 CPS 提供更为精确的抽样框。

审核内容的一些例子有:当名录上的单元数目超过或低于预定数目时,是否有恰当的说明,是否有新名单的出现,是否发现了多余的单元、是否有名单上漏掉的单元,是否有样本单元没有序列号的情况,单元名称是否有变化,是否有样本单元被破坏或废弃

(三) 样本登记

样本登记描述了一个工作过程。调查组织部门凭此来确认访问员是否找到全部的样本单元。由于 CPS 全部借助电脑进行访问,因此可以对样本实施追踪,达到对样本进行更密切的控制。样本登记主要被用来控制和检验样本的数量,这也有助于访问员工作任务的平衡分配。

三、无回答误差及控制

CPS 中有许多因素导致无回答。一种情况是被调查住户为空户,或尚未建成,或已毁损,或为非目标居民(如外国人)等。CPS 不把这种情况视为无回答,因为它们本来就在调查范围之外。这里所说的无回答指拒绝回答、没有能力回答、或由于其他原因(如外出)而无法取得联系等。在 CPS 中,把这类无回答称为 A 类无回答。

无回答还有其他一些类型。住户中的某个人可能拒绝接受访问,从而引起个人无回答。个人无回答在 CPS 中不是主要问题,因为住户中的其他人可以代其回答。另外一种项目无回答,即被调查者接受调查,但拒绝回答某个或某些特定问题。在 CPS 中,有处理项目无回答的插补程序,但这并不能保证消灭项目无回答所带来的误差,项目无回答还可能将潜在的各种偏差引入估计。

对无回答进行控制的主要措施有以下几方面。

(一) 现场指导

CPS 有专门的文件,指导现场访问员的操作。结合其他有关的信息来源,管理人员可以对现场访问员的表现做出评价。如果一个访问员的 A 类无回答率,或者在每个样本单元花费的平均时间超过正常工作水平的 1/4,该访问员就需要接受额外的培训。全国和地区在回答率方面的资料还被用于地区部门对自己范围内的现场操作情况的评价,以决定是否采取其他补充措施。

(二) 概况统计表

对无回答进行监控的另一个措施是概况统计表,它由调查总部制作,用来了解回答类型和回答方式的变化,并被用于评估数据的质量。概况统计表包括的内容大致有:地区的无回答率;与前一年月度相比的情况;无回答变为回答的转化率;计算机辅助电话调查样本的比较;每月新样本的访问情况等。

(三) 专项检查

在一份“方法与表现评估备忘录”(Reeder 1997 “Regional Response to Questions on CPS Type A Rates”, Bureau of the Census, CPS Office Memorandum No. 97-07, Methods and Performance Evaluation Memorandum No. 97-03, January 31, 1997) 中详细记载了联邦普查局和劳工统计局组成的工作小组对 CPS 中无回答状况进行检查的情况。检查的目的之一是寻找 CPS 回答率下降的原因和解决办法。检查中提出了 31 个问题,以便了解 CPS 的现场操作,并了解实施部门对无回答率上升原因的看法,以及对无回答现象进行控制的措施。该项工作得到了如下信息:(1) 访问员能够尽快地让地区部门了解无回答住户的情况。(2) 在大部分地区中,对无回答出现后所采取的补救措施有文字说明。(3) 大部分地区对特定情况下的明确拒访采取了使其转变的补救。(4) 所有地区对访问员提交的月度无回答报告都给予了及时的信息反馈。(5) 有约半数的地区对访问员进行了特殊的技术培训,以处理调查中的无回答问题。

由此看出,CPS 对控制调查中的无回答有一套比较完整的程序,这些程序包括:实施调查的操作性手册,使访问员有章可循;调查结果的汇报制度,使有关人员及时掌握无回答的情况及变动趋势,以便有针对性地采取补救措施;定期的专项检查和问题研讨;对访问员进行专门的技术培训。

四、回答误差及控制

回答误差指受访者向调整人员提供的答案与真实情况不一致。在 CPS 中,产生回答误差的原因主要有以下一些:(1) 理解问题。受访者对问题理解有偏误,因而提供了错误的答案。(2) 记性问题。由于记忆不清或不知真实答案而猜测,任选其中一个答案。(3) 心理问题。由于某些原因,有意夸大或缩小某项回答。(4) 访问员问题。由于访问员念错,或没有遵守规则的跳问,导致对方回答错误,也包括访问员记录答案的错误。

在 CPS 中,对回答误差的主要检控措施有以下几方面。

(一) 软件

20 世纪 90 年代以后,计算机辅助调查在 CPS 中发挥了重要作用。现在的软件

技术能够为每次访问提供经过自动选择的问题。计算机屏幕显示出答案的选项,访问员无须再担心由于跳问不当而导致出错。题项的文本中会自动填入规范的名称、代词、动词和参考日期。如果在提某个问题时受访者拒答,那么以后则不再对该项目提问,空项目由项目无回答的插补程序去处理,这样就把由于回答失真所带来的非抽样误差与无回答带来的误差区分开来。软件还可以在调查实施中进行逻辑审核,使访问员有机会对不正确或不一致的信息做出判断和纠正。

(二) 问卷

现在CPS的问卷是不断修订的结果,修订的目的之一是减少由于问卷—受访者—访问员之间的相互影响而导致潜在的应答误差,提高有关概念的可测量性。具体的方法包括:更短更清楚的问题措辞;把复杂问题拆分为两个或更多的问题;问题的措辞中表现出对概念的定义;对受访者主动提供信息的依赖要减少;采用不同的策略使受访者提供数字信息;对开放性问题实行预编码等。

(三) 检验与改进

研究人员在对CPS设计方案的不断修订和对调查问卷的不断研究中积累了许多经验,这些经验又被用于对另外一些新问题的研究,以保证这些新问题不会给采集数据带来失真的危险,同时这些问题又是调查内容所需要的恰当问题。

对现有调查问题进行改进也很有价值。对某些回答“不知道”或拒答比例较高的问题进行专门的分析,对访问员的调查总结进行复核。访问员和督导之间的小组会议也定期举行。

尽管对老问题和旧方法的改进将有益于CPS的数据质量,但仍需要对这些改进进行试验和效果评估后才能正式实施。例如,在引入CPS重新设计方案的头5个月,同时也在实施一项仍使用原方案的平行调查。平行调查的结果用于对新方案在无回答率和数据估计等方面产生的影响进行评估。

(四) 对访问员的培训和监督

对访问员进行集体培训和个别指导是每个地区部门为控制各种非抽样误差都要进行的连续性工作。培训内容包括增加访员的责任心、严格按程序操作、学习处理疑难问题的技巧。

现场监督是用来检查和改进现场访问员工作的方法之一,它为评价访问员的工作态度和电脑使用情况提供了一个正式渠道。共有三种监督:最初工作的监督、一般表现监督和特别需要监督。在所有这些监督过程中,督导都会强调良好访问技巧的保持、按要求的措辞提问、遵守调查手册的规定、知道怎样进行问题探索、进行详细的调查记录、在受访者改变已经给出的信息时做出正确的判断和恰当的处理、确定进行访问的最适当时间和场合等。

(五) 代答效果评价

现场访问要求每个受访者提供关于自己的信息,但真正实施是很难的。从时间和效率的角度出发,具有足够知识的成年住户成员可以为本住户其他成员代答。所以在 CPS 资料中,大约有一半的数据来自于代答。非住户成员的代答只有在特殊的情况下才被允许。

有大量的研究对受访者自答和代答的效果进行了评价。许多研究指出,自答比代答更可靠,尤其在可能有动机上的原因使二者发生差别时更是如此。例如,对于孩子的情况,父母倾向于更好的描绘。但在某些情况下代答的回答可能更准确,如对某些敏感性问题的回答。

(六) 访问员与受访者的互动

与受访者形成互动关系是提高数据质量的一个有效手段,对新样本的每第 1 个月和第 5 个月的面访更需要注意(第 1,5 次调查为面访,第 2,3,4,6,7,8 次为电话调查)。通过表现出对受访者真诚的理解与兴趣,可以建立一种友好的气氛,受访者就愿意诚实、公开地回答问题。访问员要精确地按规定的措辞提出每一个问题,如果对方没有理解或误解,就将问题重复一遍。如果仍未获得需要的回答,就采用一些探索性技巧。要创造一个良好的氛围,以利于以后的接触。

(七) 再访问程序

对一些住户进行两次调查,将两次调查的结果进行比较,对其中的差异进行分析。实际上这项工作也对指导手册、培训工作和工作程序同时进行了评价。



附录 1

方差估计软件的介绍与比较

几乎所有的统计分析软件都可以计算简单随机抽样的方差,并能够进行加权估计,例如 SAS 和 SPSS。但很多软件都没有考虑分层、多阶段等因素,也无法对方差估计进行加权处理,因而无法计算复杂抽样设计的抽样方差。对于实际抽样中的方差估计,一般需要有专门的方差分析软件。附录 1 介绍方差估计的几种主流软件并对之进行比较,附录 2 则介绍 PC CARP 软件的基本用法,以及如何应用它对缺失数据进行处理。

一、方差估计软件概述

与抽样调查的广泛应用相适应,抽样方差估计的软件发展也十分迅速。美国统计学会(ASA)的调查研究方法分会甚至专门建立了网页介绍调查分析软件,该网址为:

<http://www.fas.harvard.edu/~stats/surveysoft/survey-soft.html>

通过该网页可以了解方差分析软件的一些最新动态,网上所列出的主要调查分析软件有:

- **Bascula** 来自 Statistics Netherlands(荷兰统计局)。
- **CENVAR** 来自 U.S. Bureau of the Census(美国普查局)。

- **CLUSTERS** 来自 University of Essex(Essex 大学)。
- **Epi Info** 来自 Centers for Disease Control(疾病控制中心)。
- **Generalized Estimation System (GES)** 来自 Statistics Canada(加拿大统计局)。
- **IVFware** (beta version) 来自 University of Michigan(密歇根大学)。
- **PCCARP** 来自 Iowa State University(艾奥瓦州立大学)。
- **SAS/STAT** 来自 SAS Institute(SAS 研究所)。
- **Stata** 来自 Stata Corporation(Stata 公司)。
- **SUDAAN** 来自 Research Triangle Institute(三角研究所)。
- **VPLX** 来自 U. S. Bureau of the Census(美国普查局)。
- **WesVar** 来自 Westat, Inc(Westat 公司)。

二、四种方差估计软件简介

以上抽样方差软件都可以进行复杂样本的方差估计,但不同软件采用的估计方法、功能和特性有很大差异。在选择软件前,有必要了解这些软件的特征和性质。这里主要介绍 PC CARP, Stata, SUDAAN 和 WesVar 等四种主流方差估计软件。

1. PC CARP

PC CARP 软件 1986 年由艾奥瓦州立大学统计实验室(Iowa State University Statistical Laboratory)研制。

(1) 适用的抽样设计: 专为多阶段分层抽样设计, 对两阶段抽样可计算有限总体校正系数。

(2) 提供的估计和统计分析: 主要用于构造总量、均值、比率和比率差异的估计值, 计算估计量的标准差、变异系数、设计效应等, 还可进行加权回归估计。

(3) 方差估计主要方法: 泰勒线性近似法(Taylor linearization)。

(4) 运行环境: DOS 或 Windows 操作系统。

(5) 数据要求: 文本数据文件, 数据按初级抽样单元排序, 对数据文件还要建立一个相对应的变量名文件, 其格式为只包含一列变量名的文本文件。对观测值数量没有限制, 最多可同时计算 50 个变量的方差。

(6) 软件的一般性描述: 有文本菜单屏幕的单独程序, 数据作为 ASCII 文本输入处理

(7) 价格: 基本模块价格 300 美元, 若增加 logistic 回归和事后分层功能, 每个模块增加 50 美元, 操作手册每份售价 8 美元。

(8) 联系信息:

Sandie Smith, Statistical Laboratory, 219 Snedecor Hall, Iowa State University
Ames, IA 50011 1210
Phone: (515)294 9773
Fax: (515)294 - 4433
E mail: sandie@iastate.edu
Web page: <http://www.statlab.iastate.edu/survey/index6.html>

2. Stata

Stata 软件由美国 Stata 公司研制。

(1) 适用的抽样设计:适用于分层抽样、整群抽样、多阶段抽样的方差估计,对层内样本单位的不放回简单随机抽样可计算有限总体校正系数。

(2) 提供的估计和统计分析:包括均值、总量、比率、比例、线性回归、Logistic 回归和 probit;并提供点估计、相应的标准差、置信区间、全总体和子总体的设计效应;还可以对估计量的线性组合提供以上所有信息,并进行假设检验。

(3) 方差估计主要方法:泰勒线性近似法(Taylor linearization)。

(4) 运行环境:Windows 95, WindowsNT, Windows 3.1, DOS 等。

(5) 数据要求:最大观测值数量仅受计算机内存的限制,最大变量数为2 047

(6) 软件的一般性描述:Stata 是一个具有全面的统计功能、数据管理和图形功能的统计分析软件包。它可以交互式或批式运行,可完全编程。调查命令是标准软件包的一部分。软件可以直接读取 ASCII 文件和 Stata 格式的数据文件,其他文件形式的数据则可以通过一个独立软件包转换为 Stata 格式。

(7) 价格:一次性购买、升级购买可选。优惠措施:学术机构优惠;全套优惠;学生优惠。例如学术机构购买一个单用户版本价格为 395 美元(含有关资料)。

(8) 联系信息:

Stata Corporation, 702 University Drive East, College Station TX 77840
800 - 782 - 8272 (U.S.), 800 - 248 8272 (Canada), 409 - 696 4600
(Worldwide)

Fax: 409 - 696 - 4601

E - mail: stata@stata.com

Web site: <http://www.stata.com>

3. SUDAAN

SUDAAN 软件由美国三角研究所(Research Triangle Institute)研制。

(1) 适用的抽样设计:适用于分层样本、整群样本或多阶段样本的数据。适用于不等概率样本数据、有放回样本或不放回样本。对于任意层和任意阶样本都可进行

分析。此外,还适用于对同一总体的不同部分采用不同抽样方法的设计。

(2) 提供的估计和统计分析:包括 MULTILOG(多元 logistic 回归)、REGRESS(回归)、LOGISTIC(logistic 回归)、SURVIVAL(生存分析)、CROSSTAB(列联分析)、DESCRIPT(描述统计)、RATIO(比)。此外,EFFECT语句使用户可以进行回归系数的对比,以及单效应假设检验。

(3) 方差估计主要方法:结合使用泰勒级数线性化方法(对回归模型的 GEE)和适合于抽样设计的方差估计公式。由于抽样设计能被程序直接指定,所以用户无需编制特殊的复制权数(replicate weights)。也支持刀切法(Jackknife)和平衡半样本方差估计方法(BRR)。

(4) 运行环境:Windows 3.1,MS-DOS,Windows 95, Windows NT 以及 OS/2。

(5) 数据要求:对变量数和观测值数量都无限制。

(6) 软件的一般性描述:SUDAAN 用的是类似 SAS 的语句。在 Sun/Solaris、Windows 95 或 NT 平台下,SUDAAN 可作为 SAS 的一个程序直接调出。在其他平台下 SUDAAN 也能读 SAS 文件和读 SPSS 文件。

(7) 价格:对于 PC 机用户,SUDAAN 既可以分年度购买许可证,也可一次性购买。学术机构 PC 机的许可证年度费为每用户 50 美元 ~ 300 美元。大学的 PC 机许可证可免费发给学生用于学习目的。政府和商业机构的 PC 机许可证年度费为每用户 50 美元 ~ 450 美元。一次性购买需要 995 美元,以后升级会有优惠。学生可以用 295 美元购买一个两年的许可证。在大型机和工作站上使用 SUDAAN,享受学术机构的优惠。

(8) 联系信息:

SUDAAN Product Coordinator, Research Triangle Institute, 3040 Cornwallis Road, Research Triangle Park NC 27709 2194

Telephone: 919 - 541 6602

FAX: 919 - 541 - 7431

E mail: SUDAAN@rti.org

URL: <http://www.rti.org/patents/sudaan/sudaan.html>

4. WesVar

WesVar 软件由 WesVar 公司研制。

(1) 适用的抽样设计:该软件较灵活,适用于分层抽样和多阶段抽样,并考虑了有限总体校正因子。只要用户给出复制权数就能估计标准差。如果复制权数由程序自身产生,外部控制变量可用于实施事后分层和无回答加权调整。此外,还可以分析多重插补的数据集。

(2) 提供的估计和统计分析:多维表的估计(最多 8 维),包括总量、均值、百分比、独立性检验,以及用户指定变量函数或表单元估计。中位数和其他百分数的估计。回归分析,包括线性回归和 logistic 回归,以及方差分析。参数估计和假设检验。

(3) 方差估计的主要方法:平衡半样本法、刀切法和其他样本复制法(如自助法)

(4) 运行环境:Windows 95, 98 或 NT。

(5) 数据要求:对变量数和观测值数量都无限制。

(6) 软件的一般性描述:软件在 Windows 环境下运行,用户可通过鼠标指明自己的要求 不需要编程。

(7) 价格:WesVar4.0 版对不同平台、不同类购买者、不同的以往许可证的售价不同,单用户售价为 350 美元 ~ 495 美元,多用户售价为 2 000 美元 ~ 3 000 美元。功能有限的 WesVar2 12 版免费,并可从因特网下载。Demo 版的 WesVar 也可免费下载。学生版售价为 25 美元。

(8) 联系信息:下载或购买程序,可浏览如下网址: <http://www.westat.com/wesvar/>

联系地址:

Westat, Inc. 1650 Research Blvd Rockville, MD 20850

Attn: WesVar, RE 33F

Phone: (301) 294 - 2006

FAX: (301) 294 - 2040

E-mail: WesVar@westat.com

三、方差估计软件的比较与选择

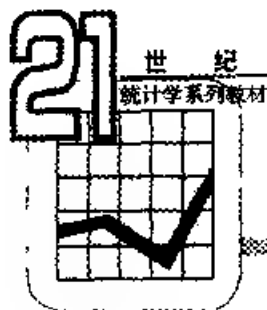
以上介绍了几种主要的方差分析软件。它们都可以进行加权、分层、多阶段抽样的点估计和方差估计,因而对每个抽样个体都要求指定权数、层数和抽样单元等,但是并非上述所有软件都可以对任意抽样设计给出无偏的方差估计。例如,只有少数软件(如 SUDAAN)可以直接处理与规模按比例的回放回抽取初级抽样单元的多阶段分层抽样设计,其他软件则需要采用最终抽样单元群模型(ultimate cluster sampling model)解决这一问题。所谓最终抽样单元群模型,就是将初级抽样单元(PSU)中的元素划分到各最终抽样单元群中,然后在初级抽样单元中无放回抽取最终抽样单元群。这样抽样设计在进行方差估计时,仅计算初级抽样单元之间的方差,却忽略了各阶段内的选择方差。采用最终抽样单元群模型的软件的方差估计都是利用这种方法。

在具体选择方差分析软件时,应主要注意以下几方面的问题。

1. 不同软件提供的估计和统计分析不同。有些软件只能计算均值、总量、比例等;有些软件还可进行 logistic 回归;还有少数软件可以计算生存分析、列联表分析、广义估计模型、特殊的比率估计的方差分析与相关检验统计量。

2. 不同软件的运行环境不同。有些软件要在 DOS 下运行,有些软件在 Windows 下运行。DOS 下运行和 Windows 下运行的软件又都可以分为两大类,即下拉菜单式(pull down menus)和命令输入式(keyword input)。比如 PC CARP 和 CLUSTERS 都是在 DOS 下运行的命令输入式软件,CENVAR 和 Epi Info 是在 DOS 下运行的下拉菜单式软件,WesVar 是在 Windows 下运行的菜单式软件,而 SUDAAN 既可以菜单操作,又可以用命令操作。

3. 不同的软件采用的复杂样本方差估计方法不同。比如 PC CARP 和 Stata 是用泰勒级数法,SUDAAN 可以使用泰勒级数法、刀切法、平衡半样本法,WesVar 可采用刀切法、平衡半样本法。



附录 2

PC CARP 软件的基本用法

这里对 PC CARP 做比较详细的介绍是因为 PC CARP 的功能比较齐全,能够满足通常调查项目的复杂样本方差估计的需要,而且该软件的价格比较便宜,操作简便,容易掌握,是国人首选的方差估计软件之一。

一、PC CARP 概述

1. 用途

PC CARP 可用来计算总量、均值、分位数、比率、比率的差以及列联表中频数的估计值和标准差,并可以进行加权回归公式的估计。该程序是为多阶段分层样本设计的,可在两个阶段引进限定修正项(finite correction term)。

2. 功能

PC CARP 的主要功能见附表 2.1,其中 \checkmark 表示可用选项。对大多数估计量都提供其估计值、估计值的标准差,以及变异系数等选项。在不同的情况下,对变量个数有一定的限制(见附表 2.1),但对于层数和每层中的群数没有限制。

附表 2.1

PC CARP 功能一览表

分析	协方差矩阵	设计效应	备注
总体分析			
总量估计	✓	✓	最多 50 个变量
比率估计	✓	✓	不需要协方差,最多 50 个变量
比率的差		✓	需要协方差,最多 15 个变量
分层分析			
总量		✓	最多 50 个变量
均值		✓	最多 50 个变量
比例		✓	最多 50 个变量
子总体分析			
总量		✓	多变量分层
均值		✓	多变量分层
比例		✓	多变量分层
比率		✓	多变量分层
其他分析			
列联表	✓		最多 50 个单元,比例检验
回归	✓		最多 50 个变量
单变量		✓	假设检验、残差、预测值 多变量、经验分布函数 分位数、分位数间距离

本程序可针对以下三种情况计算方差:第一种是可看做来自无限总体的样本。第二种是层内初级抽样单位(群)的抽样比由用户给定。在给定的抽样比下,PC CARP 对每层都可以计算带有限总体校正项的方差。第三种是针对两阶段抽样,并要求用户提供第一阶段各层的比例和权重。如果选择了两阶段选项,所用的方差估计公式在每个阶段都有有限总体校正项。此时程序将所有观测单元都视为第二阶段的单位,并且利用观测单元的权重计算第二阶段的抽样比。

PC CARP 有合并层的功能,即将只含一个群的层与排在其后的层组合成一个层,合并后将形成一个全新的数据集。

3. 数据

PC CARP 的输入文件主要是观测数据(向量)文件,其内容包括控制变量和分析变量两部分。此外还可以有一个抽样比文件,其内容就是各层的抽样比。

常见的数据源就是调查数据,数据源应具有层标识和群标识,而且每个观测单元都有相应的所有变量值,通常每个观测单元都有一个权重,该权重是被抽中概率

的倒数。PC CARP 接受的数据应是按照观测单元组织好的数据,此外还包括层数、群数和权重等控制变量。PC CARP 读入程序的数据必须按照层中的群排序。正是这些数据的排列顺序定义了数据的层次和群结构。

控制变量一般包括层标识、群标识以及权重(这里权重是被抽中概率的倒数,需输入程序)等三部分。其中,层标识必须是标识的第一部分,群标识应紧跟层标识后面,而且两者应都是整数,层标识和群标识都不能超过 10 位数字。

如果每个观测单元被抽中的概率都相同,这样的样本是自加权的,它不需要为每个观测单元输入权重。可接受的控制变量组合包括:(1) 层、群和权重(完全调查);(2) 层和群(自加权);(3) 群(群的随机抽样);(4) 无标识(简单随机抽样)。

以上组合之外的控制变量不会被程序接受。如果没有输入权重,每个观测单元都被分配一个值为 1 的权重。如果选择了两阶段选项,就必须提供权重。

如果分析中需要有限总体修正项,那么程序就要求给出第一阶段的抽样比。抽样比必须放在数据文件外的另一个文件中,并且对应每个层都要有抽样比。

数据文件的格式可以是以下两种类型:(1) 有向列表(list directed)。文件中的数字由空格或逗号分隔开。一个观测单元的数据可以延续到下一行,但每个新观测单元必须始于一个新行。(2) 格式化(formatted)。在这种情况下,PC CARP 需要一个 FORTRAN 格式。一个观测单元的数据可以延续到下一行,但每个新观测单元必须始于一个新行。如果数据存在多个文件中,每个文件必须有相同的格式。如果数据集包括层标识和(或)群标识,它们必须是整数形式(I),而且按层,随后是群的顺序排列。其他所有变量,包括权重,必须按照一种实数格式(F,D 或 G)来读取(请参见后面案例)。

4. 程序用法说明

PC CARP 采用菜单驱动,而且大多数菜单都是自解释的。第一组菜单称为“问题定义”(problem specification),主要用来定义所要分析的变量,对这些变量命名,并且将数据集及其相关标识提交给程序。菜单的第二部分称为“分析定义”(analysis specification),主要是用来选择要执行的分析类型,识别参与分析的变量,并且输入分析所需的其他指令。

问题说明阶段分为 5 个菜单,分别是问题说明、变量命名、数据类型、数据获取及数据输出。每个菜单可能有多个屏幕显示。

在问题说明的某些位置,可以输入 G 来表示“Go back”,使得用户回到先行菜单,并使得用户能够在不终止程序执行的情况下进行一些改动。需要注意的是,Go back 操作使得用户回到先行菜单,而非先前显示屏幕,因为先行菜单可能不只一个显示屏幕。

如果用户并不清楚系统要求的信息,可输入 H 来寻求帮助。

输出可以直接输出到打印机,或产生一个输出文件,或两个选项都选。输出结果可分为两个部分:第一部分是在问题定义阶段后就获得的,是由用户给定的数据的概览。第二部分包括各指定分析的统计结果。结束一次“问题定义”阶段后,可执行多次分析。输出根据分析类型的不同而不同,而且输出在很大程度上是自解释的(请参见后面案例)。

二、案例应用

以某两阶段分层调查为例,假如第一阶段按照行政区域分层,共分为 4 层,各层(strata)代表特定的行政区划,在各行政区划内抽取居委会,这里的居委会就称为群(clusters),共 14 个群;第二阶段观测单元则是居委会内的人,共 21 个。数据见附表 2.2。

1. 总数、均值、比率、比率差分

根据 PC CARP 的要求,数据已依次按照层和群排序。它的格式是(2I2, 6F4.0)。层次 1,2,3,4 的比率分别是 0.10,0.05,0.20 和 0.25。这些比率在例子的文件 rates1.dat 中。这些比率的格式是(F4.2)。

附表 2.2 案例数据

层(stratum)	群(cluster)	权数(weight)	Y2	Y3	Y4	Y5	Y6
1	1	10	10	11	2	1	1
1	2	10	11	13	2	2	2
1	3	10	12	10	1	3	1
1	4	10	8	7	2	1	1
1	5	10	6	5	1	2	1
1	5	10	4	9	1	3	2
2	1	20	3	6	1	2	1
2	2	20	6	10	2	1	1
2	3	20	14	12	2	1	1
2	4	20	6	4	1	2	1
3	1	5	12	15	3	4	1

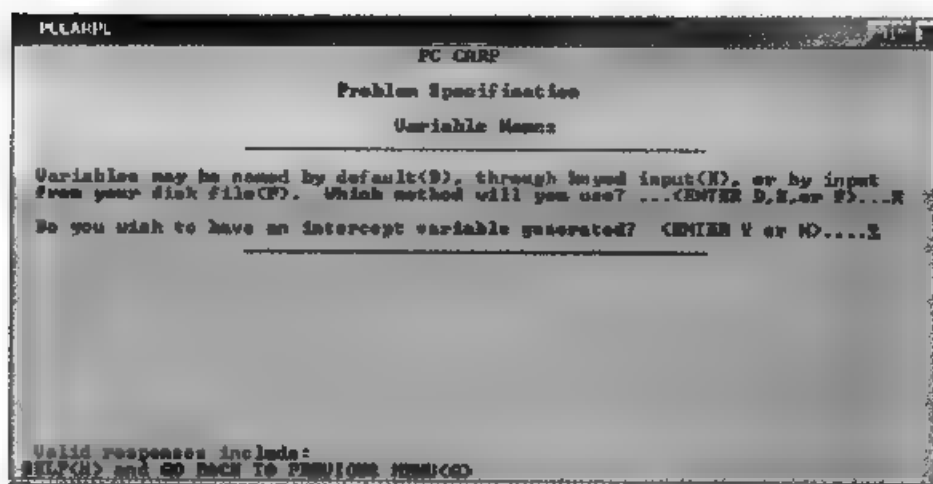
按 Enter 键进入“变量命名”屏幕,如附图 2.4 所示。变量命名有三种方式可选用:

D 为缺省值 (default), 该方法采用程序缺省变量名, VBLE01, VBLE02, ..., VBLE tn, 这里 tn 是所有输入变量的总数。

键入 K, 即键盘输入变量名, 程序将提示用户输入变量名。

键入 F, 即从磁盘文件中输入变量名, 程序将提示用户输入文件名。该文件中每个变量名应该单独一行, 每行都不能超过 8 个字符。

本例中选择用键盘输入变量名, 所以键入 K, 如附图 2.4 所示。



附图 2.5

按 Enter 键后, 如附图 2.5 所示, 程序询问是否需要截距变量 (intercept variable), 缺省回答是 Y。该变量通常是为满足分析的需要, 如果拿不定主意就先输入 Y。

本例中应分析所需, 键入 Y。在加入截距变量后, 变化后的数据集如附表 2.3 所示, 程序自动加入取值为 1 的 Intercept 变量。注意, 截距变量被放在分析变量中的第一位。

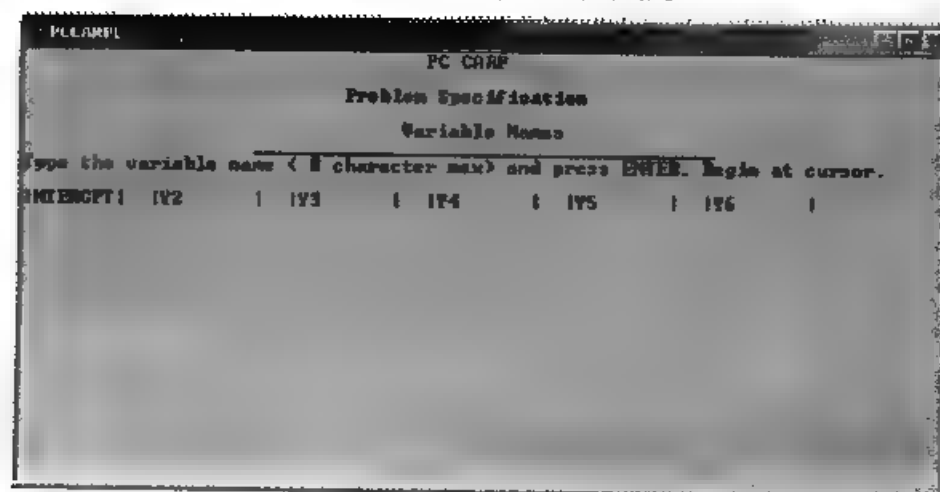
附表 2.3 加入截距变量后数据

stratum	cluster	weight	Intercept	Y2	Y3	Y4	Y5	Y6
1	1	10	1	10	11	2	1	1
1	2	10	1	11	13	2	2	2
1	3	10	1	12	10	1	3	1
1	4	10	1	8	7	2	1	1
1	5	10	1	6	5	1	2	1
1	5	10	1	4	9	1	3	2

续前表

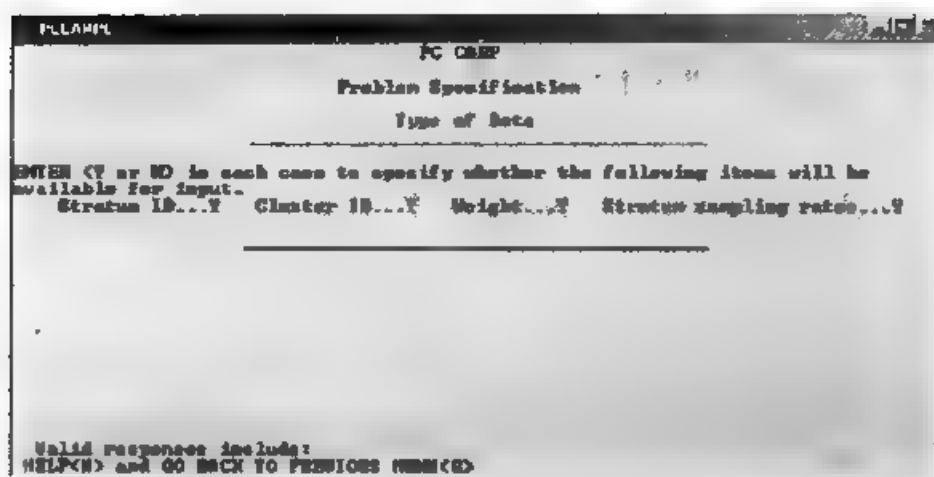
stratum	cluster	weight	Intercept	Y2	Y3	Y4	Y5	Y6
2	1	20	1	3	6	1	2	1
2	2	20	1	6	10	2	1	1
2	3	20	1	14	12	2	1	1
2	4	20	1	6	4	1	2	1
3	1	5	1	12	15	3	4	1
3	1	5	1	1	4	3	4	1
3	2	5	1	2	3	3	4	1
3	2	5	1	3	1	3	4	1
3	2	5	1	5	6	3	4	2
4	1	4	1	9	7	2	1	1
4	1	4	1	7	4	1	2	1
4	2	4	1	10	12	2	1	1
4	2	4	1	15	14	2	2	2
4	3	4	1	5	8	1	3	2
4	3	4	1	7	7	1	2	1

按 Enter 键后,就显示附图 2.6,要求输入分析变量名。INTERCEPT 总是 PC CARP 的第一位。因此,如果用数字来命名建议从数字“2”用作 Intercept 后的第一个变量的名字。每个变量名都不能多于 8 个字符。在输入一个变量名之后,按回车键会使光标移动到下一个名字的正确位置。如果不小心输入错误,可输入 G,回到上一个名字更改。附图 2.6 显示输入变量名后的屏幕。



附图 2.6

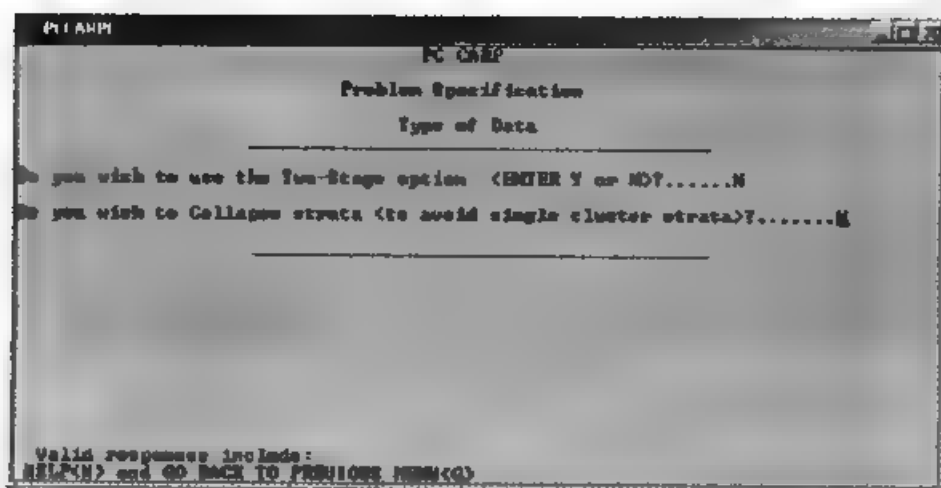
在第6个名字(也就是我们输入的第5个变量名)后按Enter键,出现附图2.7.



附图 2.7

附图 2.7 是“数据类型”菜单,用于识别抽样设计的有关信息、层标识、群标识以及权是用来识别每个观测的项目,可接受的组合有:(1) 层次、群和权重(完全调查);(2) 层次和群(自加权);(3) 群(群的随机抽样);(4) 无标识(简单随机抽样)。

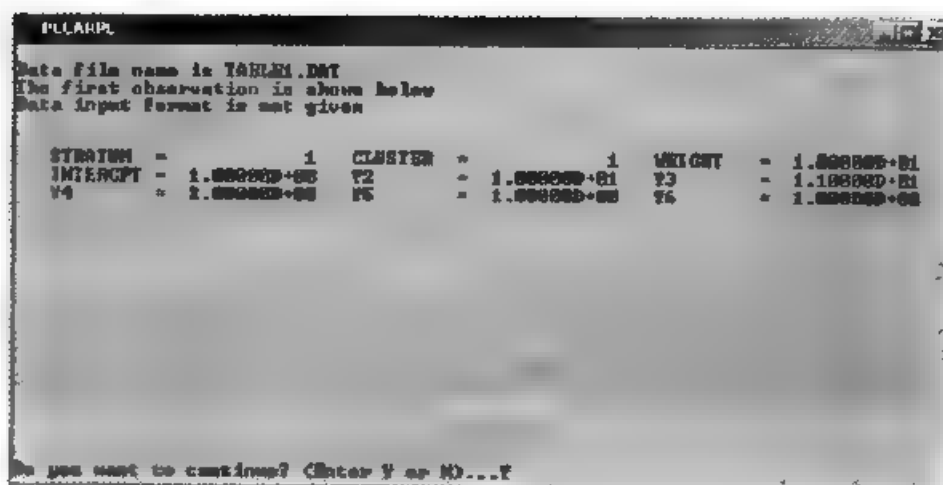
程序不接受任何其他的组合。当没有提供权重时,每个观测的权重都被设置为1。如果打算选择两阶段选项,就必须提供层标识、群标识、权重以及层抽样比。本例中使用所有的缺省值Y。因此,对四个项目都按Enter键,出现屏幕询问是否选择两阶段抽样。本例中不需要,所以输入缺省值N,见附图2.8。按Enter键后,屏幕询问是否有合并层的问题。由于本例中已确定每个层都至少含有两个群,所以不需要进行合并,输入缺省置N。



附图 2.8

按 Enter 键后,PC CARP 显示第一个观测单元的数据,以便审核有无格式错误,见附图 2.11,并询问是否继续。审核无误后,输入 Y。如果输入 N,程序将中止。本例输入 Y

按 Enter 键后,屏幕询问是否还有其他数据文件,本例无其他数据文件,输入 N。



附图 2.11

按 Enter 键后,屏幕询问抽样比文件是否为格式化形式?本例输入 Y。

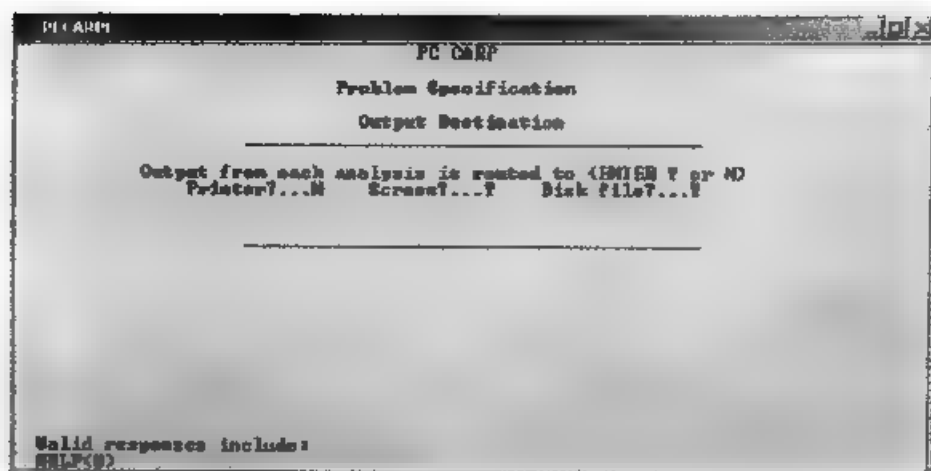
按 Enter 键后,屏幕询问抽样比数据格式,输入(F4.2)。抽样比文件中抽样比个数要对应层数,即每层都有且只有一个抽样比,见附图 2.12。如果比例数目偏多或偏少,程序都将中止。

0.10
0.05
0.20
0.25

附图 2.12 抽样比文件 RATES1.dat 文件格式

按 Enter 键后,屏幕询问抽样比文件名和路径。输入 RATES1.DAT。

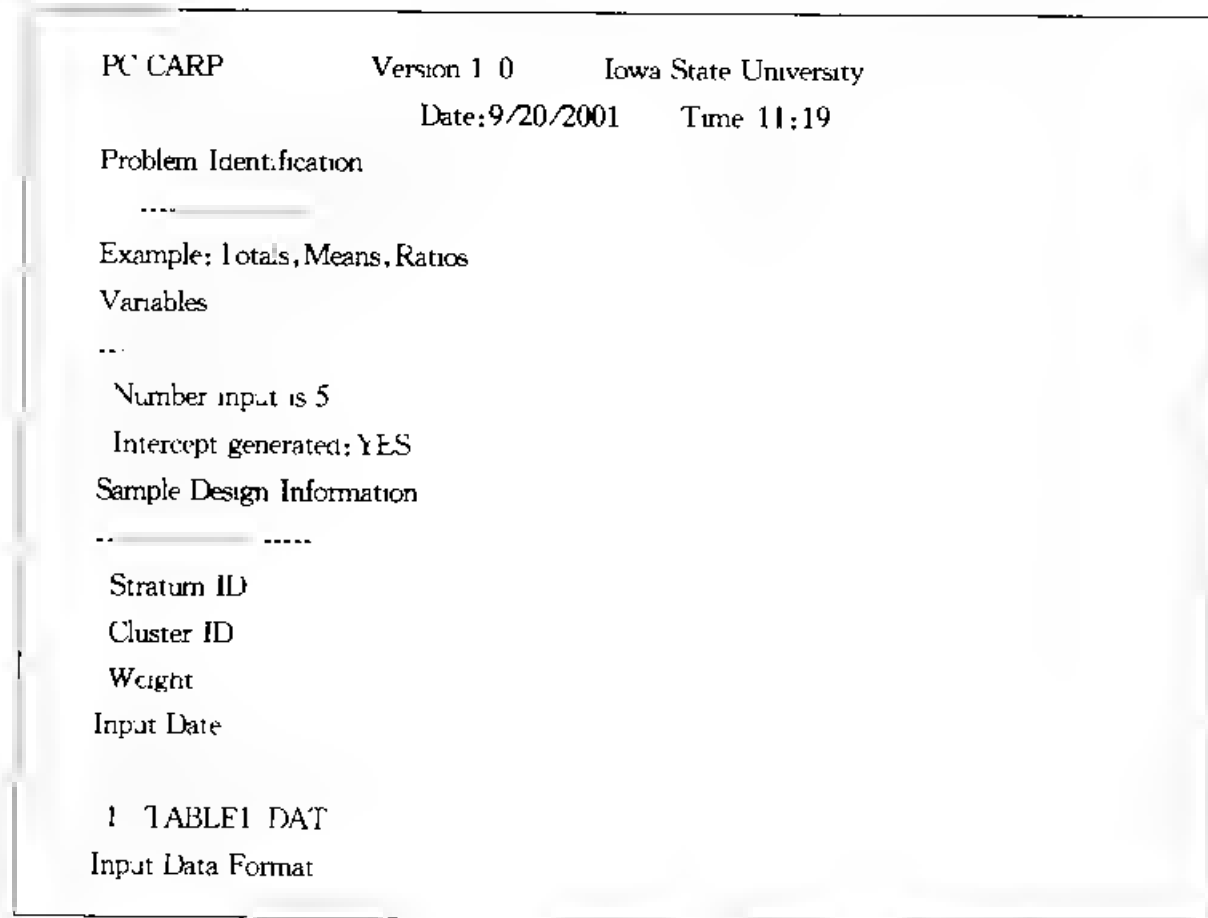
按 Enter 键后,进入“数据输出”菜单。屏幕询问输出形式,可以选择输出到打印机(PRINTER)、屏幕(SCREEN)、磁盘文件(DISK FILE)。用户在所希望的输出选项后输入 Y,在不需要的输出选项后输入 N。要注意的是,屏幕显示非常快,显示结果也不便浏览 建议在打印和磁盘文件中至少选择一种。本例中选择了屏幕和磁盘文件格式,见附图 2.13



附图 2.13

按 Enter 键后,屏幕要求询问输出文件名和路径,输入 OUTPUT.DAT。

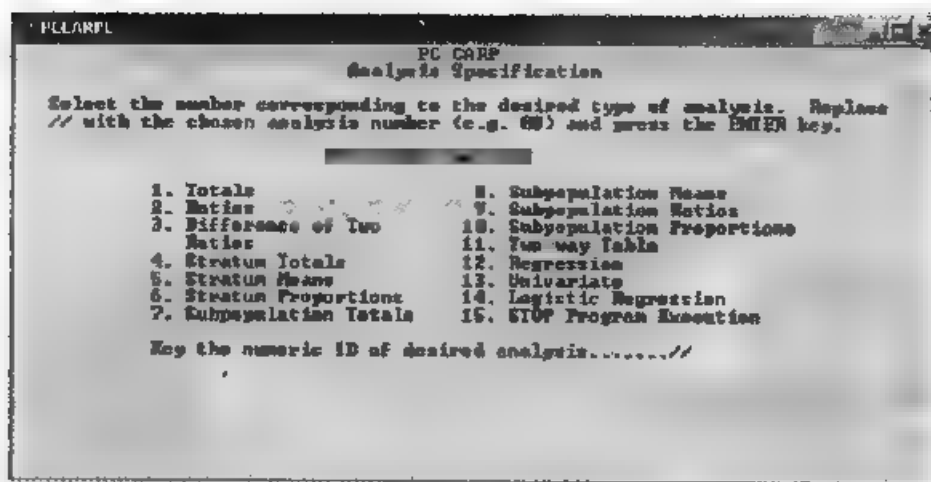
按 Enter 键后,屏幕显示对数据的定义说明,其内容见附图 2.14,至此,问题定义阶段结束,PC CARP 已获得它所需的关于数据和设计的所有信息,可以进入数据分析阶段。





附图 2.14 问题定义完成后输出的结果

“数据分析”菜单列出了各类分析选项,见附图 2.15 本例中,我们希望估计总量、比率和比率的差。

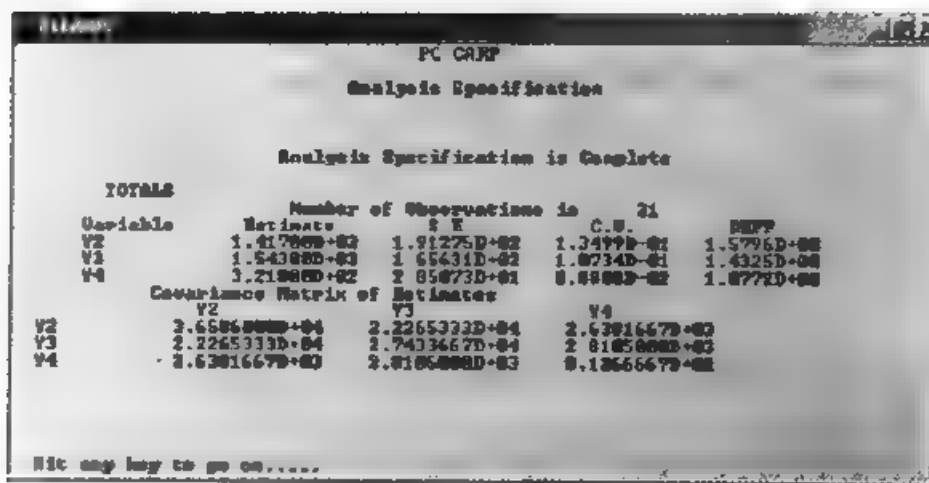


附图 2.15

首先进行总量分析,输入 01 按 Enter 键后,程序出现两个问题询问是否需要估计协方差矩阵和设计效应,本例对这两个问题都回答 Y,见附图 2.16。

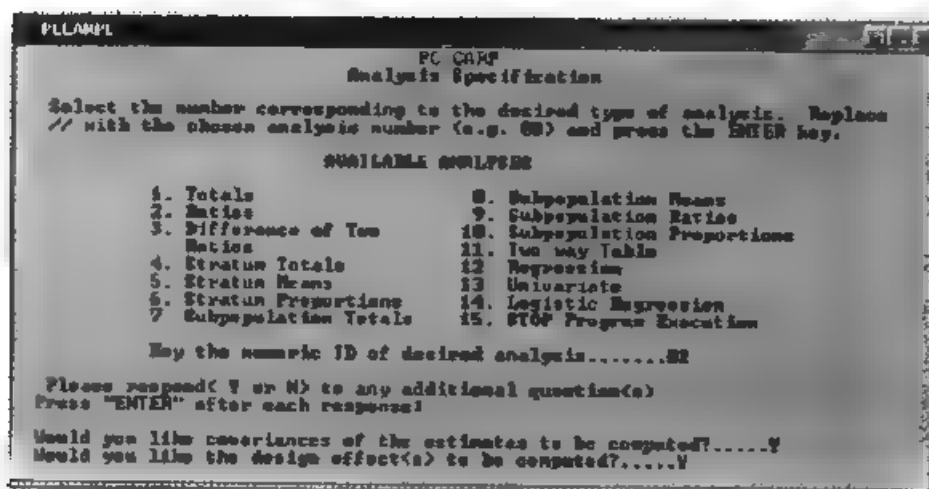
按 Enter 键后,程序要求选择所分析变量见附图 2.17。菜单上已列出可供选择的变量名,PC CARP 总是将 Intercept 变量放在首位,并指定其为第一选择变量。这里我们要对 Y2,Y3,Y4 进行总量分析,因此依次输入各变量前序号 02,03,04。

选择了所需的所有变量后,输入 Y 结束变量选择。出现菜单询问是否执行分析,见附图 2.18。如果输入 N,则返回“数据分析”菜单,重新选择;如果输入 Y,则程序开始进行分析,并输出分析结果,见附图 2.19。对应各分析变量,程序给出其总量估计(estimate)、估计的标准误(S.E.),变异系数(C.V.),设计效应(deff),以及



附图 2.19

程序返回“数据分析”菜单,我们下一步进行比率估计。输入02后,同样选择进行协方法估计和计算设计效应,见附图 2.20。

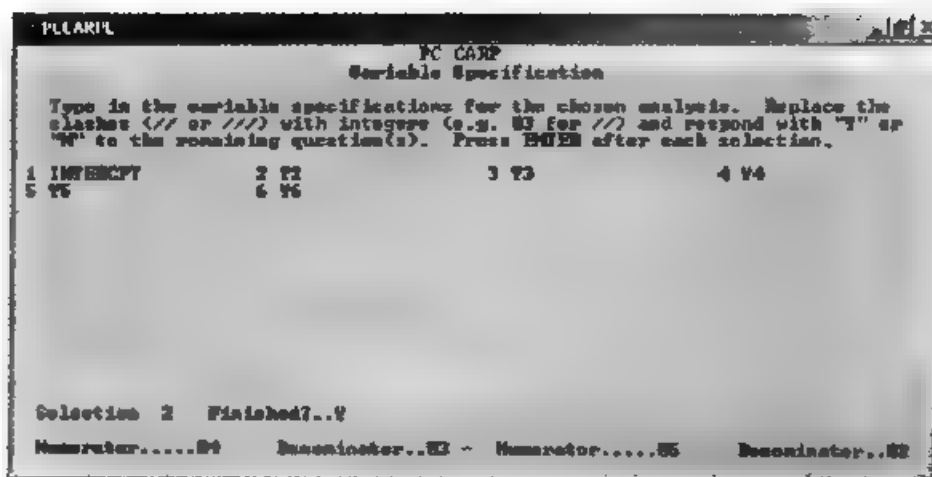


附图 2.20

按 Enter 键后,程序要求选择所分析变量,见附图 2.21。要估计一个比率,需要指定两个变量:分子变量(numerator)和分母变量(denominator)。如果分母变量指定为截距变量(intercept),比率估计也可以进行均值估计。本例中选择了4个比率,见附图 2.21,分别为,

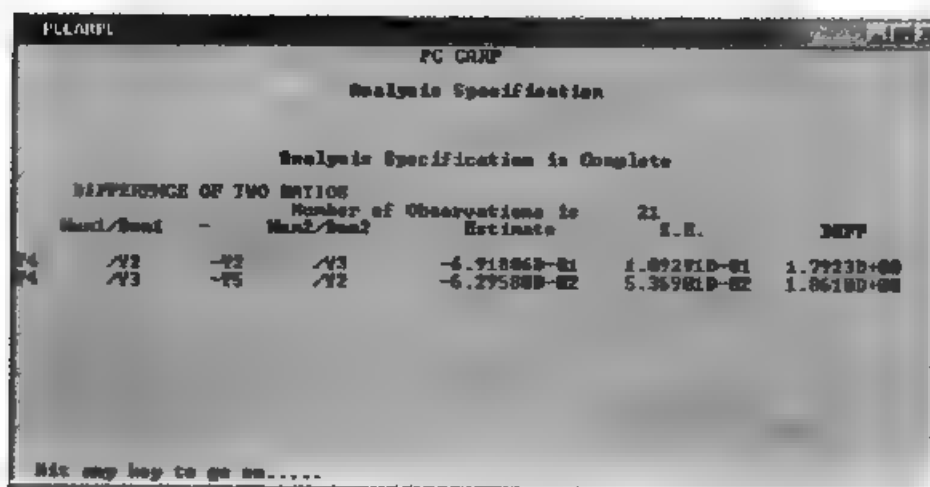
$$\frac{Y2}{\text{intercept}}, \frac{Y3}{\text{intercept}}, \frac{Y2}{Y3}, \frac{Y5}{Y3}$$

程序计算结果见附图 2.22 因为选择了4个比率,协方法矩阵为4×4矩阵 比率 $\frac{Y2}{\text{intercept}}$ 和 $\frac{Y3}{\text{intercept}}$ 的估计就相当于对 Y2 和 Y3 的均值估计,见附图 2.22 和附



附图 2.23

计算结束后又返回“数据分析”菜单,如不再继续进行分析,可输入 14,结束 PC CARP。



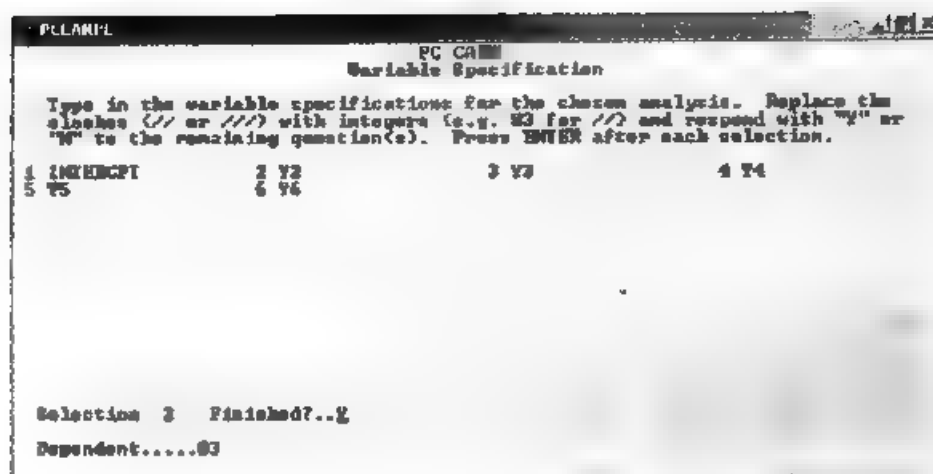
附图 2.24

2. 分层估计

利用附表 2.2 中的数据进行分层估计。关于问题定义阶段的操作同前,我们直接从“数据分析”阶段开始。在数据分析菜单中,选择 04,即分层总量分析,并输入 Y 要求计算设计效应。

进入选择变量的菜单,见附图 2.25,输入分析变量的编号 02 和 03,输入 Y 结束选择后输出结果,见附图 2.26。

分析结果显示见附图 2.26。各层内,对应各分析变量,程序给出其总量估计 (estimate),估计的标准误 (S E),变异系数 (C V),设计效应 (deff) 第二层中为简单随机抽样,故 deff 都为 1



附图 2.25

除分层总量估计外,用户还可以在“数据分析”菜单中选择 05 或 06,进行层内均值或比例估计。但 PC CARP 无法直接进行层内比率估计,如果希望进行层内比率估计,可以将层标识复制为一个实数分析变量,然后利用该变量作为分层变量进行子总体的比率分析。

PC CARP
Analysis Specification is Complete

STRATUM TOTALS					
TOTALS FOR STRATUM 1					
Variable	Estimate	S.E.	C.U.	DEFF	
Y2	5.10000E+02	3.14443E+01	6.16950E-02	1.72780E-01	
Y3	5.58000E+02	5.00000E+01	1.95630E-01	7.65310E-01	
TOTALS FOR STRATUM 2					
Variable	Estimate	S.E.	C.U.	DEFF	
Y2	5.00000E+02	1.03703E+02	3.17870E-01	1.00000E+00	
Y3	6.40000E+02	1.42361E+02	2.22440E-01	1.00000E+00	
TOTALS FOR STRATUM 3					
Variable	Estimate	S.E.	C.U.	DEFF	
Y2	1.15000E+02	1.74164E+01	1.16660E-01	9.32640E-02	
Y3	1.45000E+02	4.02472E+01	2.77500E-01	5.45450E-01	
TOTALS FOR STRATUM 4					
Variable	Estimate	S.E.	C.U.	DEFF	
Y2	2.12000E+02	3.97500E+01	1.88440E-01	1.82190E+00	
Y3	3.00000E+02	4.66047E+01	2.24060E-01	3.24000E+00	

Hit any key to go on.....

附图 2.26

3. 子总体估计

数据格式见附表 2.4,该数据只是在附表 2.2 数据的基础上增加了 2 个变量。假设后 5 个变量为分层变量,前两个变量为因变量,也称为分析变量。要注意,分层变量取值不能为负,分析变量的取值可以取负值。

该数据也是附在 PC CARP 程序中的,该文件名为 TABLE3.DAT。数据格式为 (2I2,8F4.0),各层次 1,2,3,4 的抽样比仍为是 0.10,0.05,0.20 和 0.25,该文件仍为 rates1.dat。这些比率的格式是 (F4.2)。

关于数据的定义过程同前,不再赘述,问题定义结果见附图 2.27。

PC CARP	Version 1.0	Iowa State University
	Date:9/20/2001	Time:13:50
Problem Identification		
Example: SUBPOPULATION ESTIMATES		
Variables		
Number input is 7		
Intercept generated: YES		
Sample Design Information		
Stratum ID		
Cluster ID		
Weight		
Input Date		

1 TABLE3.DAT		
Input Data Format		

List directed		
(2I2,8F4.0)		
Sampling Rates		
RATES1.DAT		
(F4.2)		
Output to disk file: OUTPUT.DAT		

附图 2.27 问题定义完成后输出结果

问题定义结束后,进入数据分析阶段。进入“数据分析”菜单后,输入08选择子总体均值估计,同样选择计算设计效应。

进入变量选择的菜单中,程序首先要求定义分层变量,见附图 2.28。本例计划选用C4和C5交叉划分总体。这里,截距变量(intercept)也可以作为分类变量,如仅选择截距变量为分类变量,则其结果等同于选择总体均值分析。

首先定义第一个分层变量(见附图 2.28):

PC CRAP
Variable Specification

Type in the variable specifications for the chosen analysis. Replace the slashes (// or ///) with integers (e.g. 03 for //) and respond with "Y" or "N" to the remaining question(s). Press ENTER after each selection.

1 INTERCPT	2 V2	3 V2	4 C4
5 C5	6 C5	7 C7	8 C8

Classification Variable Selection 1

Variable...04 Crossed with previous variable?...N

Upper bound on categories...002 Last Class Variable?...N

附图 2.28

输入变量编号 04;

因是第一个分类变量,不与上一个变量交叉,“Crossed with previous variable?”回答 N,该变量最大层数为 3;

还有其他分层变量,“Last Class Variable”,故输入 N。

然后定义第二个分层变量,见附图 2.29:

PC CRAP
Variable Specification

Type in the variable specifications for the chosen analysis. Replace the slashes (// or ///) with integers (e.g. 03 for //) and respond with "Y" or "N" to the remaining question(s). Press ENTER after each selection.

1 INTERCPT	2 V2	3 V2	4 C4
5 C5	6 C5	7 C7	8 C8

Classification Variable Selection 2

Variable...05 Crossed with previous variable?...Y

Upper bound on categories...004 Last Class Variable?...Y

附图 2.29

输入变量编号 05;

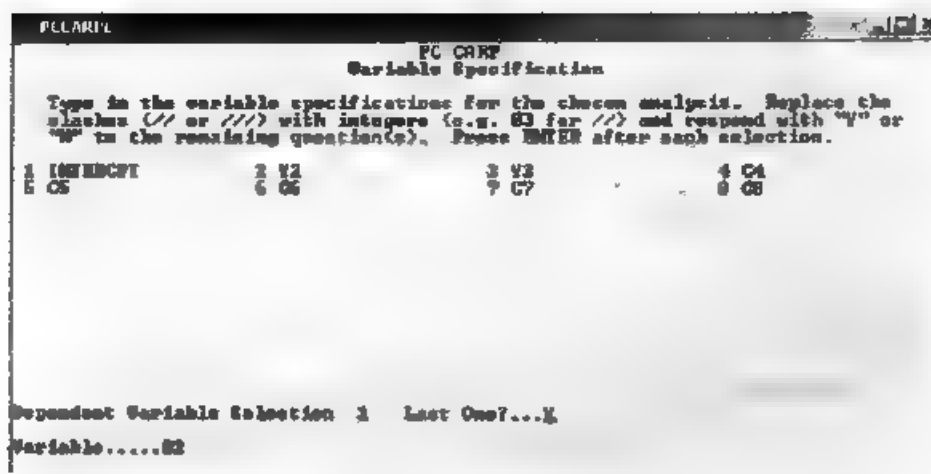
与上一个变量交叉,“Crossed with previous variable?”回答 Y;

该变量最大层数为 4;

不再输入其他分层变量,“Last Class Variable”,故输入 Y。

然后按照程序的要求选择分析变量,见附图 2.30。

程序分析结果见附图 2.31。由于观察单元有限,许多子总体内没有观察单元,估计值为空。



附图 2.30

SUBPOPULATION MEANS

Dependent variable is Y2

Category	Estimate	S E	C.V.	DEFF
----------	----------	-----	------	------

Number of observations in subpopulation is 0

C4	1.000 0	C5	1.000 0
----	---------	----	---------

* * * * *

Number of observations in subpopulation is 5

C4	1.000 0	C5	2.000 0
----	---------	----	---------

5.103 45D + 00	8.757 55D + 01	1.716 0D + 01	2.172 5D + 00
----------------	----------------	---------------	---------------

Number of observations in subpopulation is 3

C4	1.000 0	C5	3.000 0
----	---------	----	---------

7.500 00D + 00	2.537 48D + 00	3.383 3D + 01	1.261 5D + 00
----------------	----------------	---------------	---------------

Number of observations in subpopulation is 0

C4	1.000 0	C5	4.000 0
----	---------	----	---------

* * * * *

Number of observations in subpopulation is 6

C4	2.000 0	C5	1.000 0
----	---------	----	---------

9.647 06D + 00	1.892 85D + 00	1.962 1D + 01	2.919 6D + 00
----------------	----------------	---------------	---------------

Number of observations in subpopulation is 2

C4	2.000 0	C5	2.000 0
----	---------	----	---------

1.214 29D + 01	1.048 59D + 00	8.635 5D + 02	5.612 2D + 01
----------------	----------------	---------------	---------------

```

Number of observations in subpopulation is 0
C4      -      2 000 0  C5      -      3 000 0
                      * * * * *
Number of observations in subpopulation is 0
C4      2 000 0  C5      4 000 0
                      * * * * *
Number of observations in subpopulation is 0
C4      -      3 000 0  C5      1 000 0
                      * * * * *
Number of observations in subpopulation is 0
C4      3 000 0  C5      2 000 0
                      * * * * *
Number of observations in subpopulation is 0
C4      3 000 0  C5      3 000 0
                      * * * * *
Number of observations in subpopulation is 0
C4      -      3 000 0  C5      4 000 0
                      4 600 00D + 00  1.359 53D + 00  2 955 5D-01  3.562 8D-01

```

附图 2.31 子总体分析均值估计结果

三、PC CARP 软件对缺失数据的处理

1. PRE CARP 概述

PC CARP 只能对完整的数据集进行处理,但实际的调查数据中经常出现数据缺失现象。对这种不完全数据集,PC CARP 提供了 PRE CARP 程序对缺失值进行处理。PRE CARP 可以用热卡插补法(hot deck imputation)对缺失数据进行插补,这样数据集就变为可处理的完整数据集。

PRE CARP 使用的热卡插补法是用缺失值的前一个回答值替代缺失值,因此数据的排列顺序对缺失值的替代数据影响非常大。在进行插补前,一定要对数据进行合理排序。当然,这种排序的数据集不一定满足 PC CARP 的按层内群排序的要求,因此在输入 PC CARP 前可能还需要再次排序。

HOTDECK 法可以用于总数据的插补,也可用于层内插补。其原理是在各层内部用缺失值的前一个回答值替代缺失值。一般分类变量的层数不要超过 10,不论

是对分层插补还是总插补,PRE CARP 最终都要记录插补的缺失值数量。

PC CARP 对插补后的数据和一般原始数据的方差估计方法相同。要考虑插补对方差估计的影响,有一个粗略的处理方法就是将 PC CARP 估计的方差乘以一个调整系数:

$$(n_0 + n_M)^{-1}(n_0 + 3n_M)$$

式中, n_0 为观察值数量; n_M 为缺失值数量,假设 $n_0 \geq n_M$ 。

比如对于简单随机抽样,对完整数据集的方差估计公式为:

$$(n_0 + n_M)^{-1}s^2$$

由于该方差估计是基于 $n_0 + n_M$ 个数据,其中包括 n_M 个缺失数量的插补值。因此需要对方差估计结果进行调整,最终方差估计会变为:

$$(n_0 + n_M)^{-2}(n_0 + 2n_M + n_M)s^2$$

此外,PRE CARP 可以对所有带缺失值的原始变量都生成一个指示变量,如果变量值为观察数据,则其取值为 1;如果变量值为缺失插补值,则取值为零。如果将该指数变量作为分层变量用于 PC CARP,就可以直接对原始观测值进行估计了。

2. 案例

不完全数据集见附表 2.5,数据格式为(2I2,6F4.0)。对于 Y2,缺失值的代码是 99。对于 Y3,缺失值的代码是 88。对于 C6,缺失值的代码是 M。

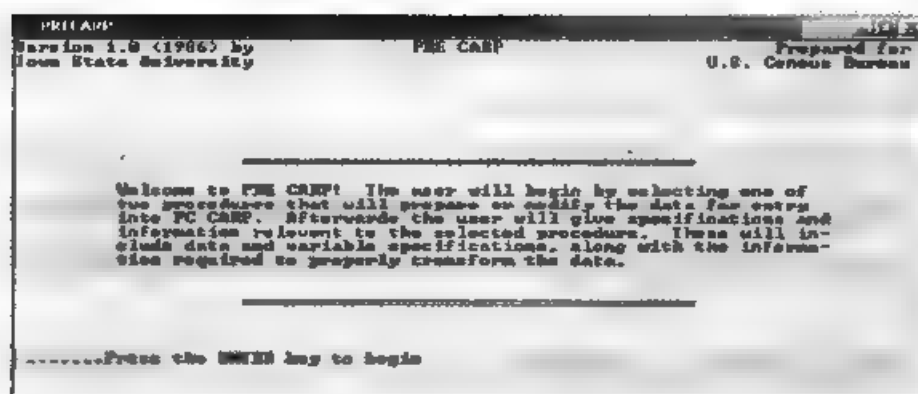
附表 2.5 缺失数据集案例数据

stratum	cluster	weight	Y2	Y3	C4	C5	C6
1	1	10	99	11	2	1	1
1	2	10	11	13	2	2	2
1	3	10	12	10	1	3	M
1	4	10	99	7	2	1	M
1	5	10	6	5	1	2	1
1	5	10	4	88	1	3	2
2	1	20	3	6	1	2	1
2	2	20	6	10	2	1	1
2	3	20	14	12	2	1	1
2	4	20	6	4	1	2	1
3	1	5	99	15	3	4	M
3	1	5	1	88	3	4	M
3	2	5	2	3	3	4	1
3	2	5	3	1	3	4	1
3	2	5	5	6	3	4	2

续前表

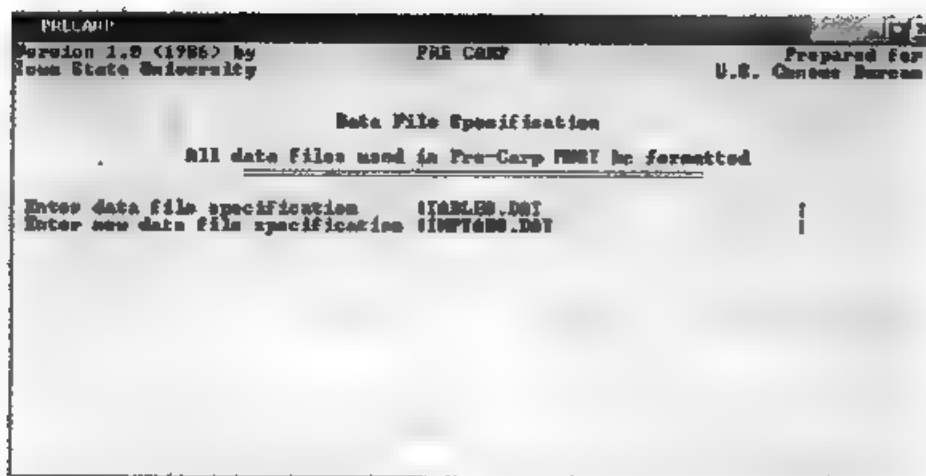
stratum	cluster	weight	Y2	Y3	C4	C5	C6
4	1	4	9	7	2	1	1
4	1	4	7	4	1	2	1
4	2	4	10	12	2	1	1
4	2	4	15	14	2	2	2
4	3	4	5	8	1	3	M
4	3	4	7	88	1	2	1

运行 PRE CARP 后,首屏幕显示是该程序的介绍,如附图 2.32 所示。



附图 2.32

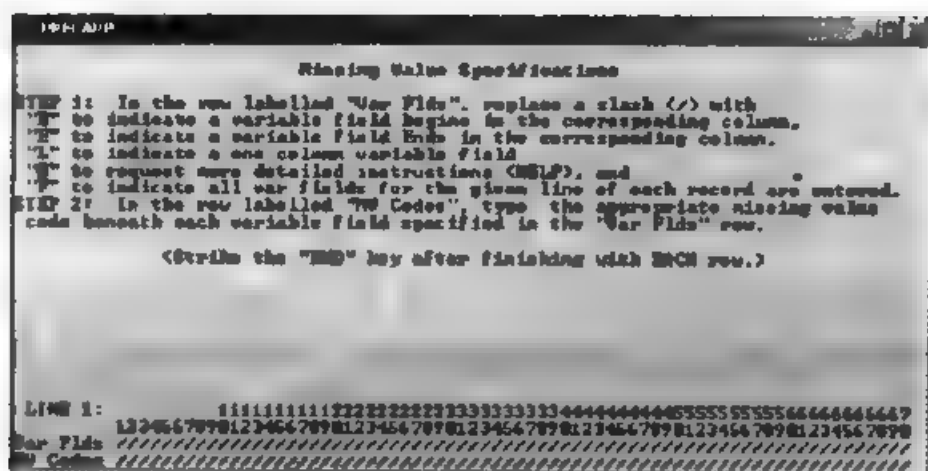
按 Enter 键进入附图 2.33,程序询问需要处理的文件名(含路径)与输出结果文件名(含路径)。



附图 2.33

按 Enter 键后,程序询问是否还有其他数据文件 本例回答 N,见附图 2.34。

附图 2.36。程序对每个含缺失值的变量都会生成一个新变量,其取值为 1,就说明相应变量值为真实观察值;取值为零,就说明相应变量值缺失。

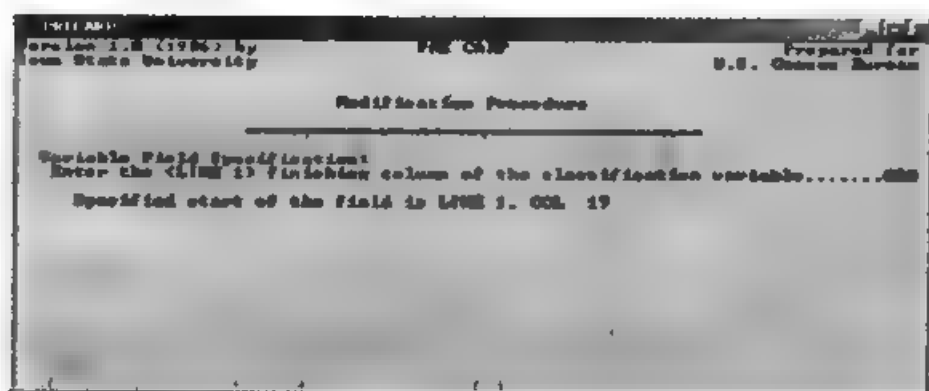


Var Flds 行是用于标明含有缺失值的变量,见附图 2.38。由于案例数据格式为 (212,6F4 0),前 8 列分别层数,群数和权重。第一个含有缺失值的变量 Y2 从第 9 列到第 12 列,因此第 9 列输入 B,第 12 列输入 E;第二个含有缺失值的变量 Y3 从第 13 列到第 16 列,因此第 13 列输入 B,第 16 列输入 E;第三个含有缺失值的变量 C3 从第 25 列到第 28 列,因此第 25 列输入 B,第 28 列输入 E。指定完缺失值的区域后,输入 F,然后按 End 键,光标移至下一行。

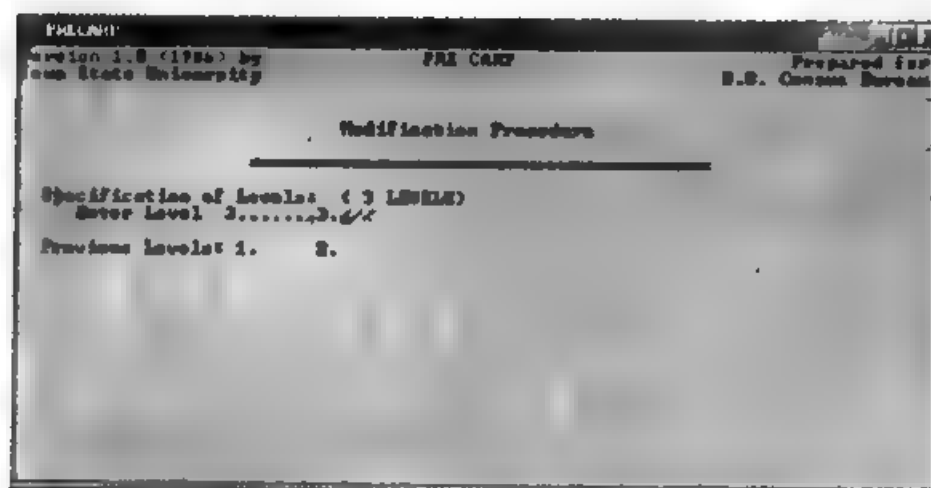
附圖 2.39

88;C3 的缺失值代码为 M,则在第 25 列到第 28 列输入 M.输入 F,然后按 End 键。程序询问每个记录覆盖的列数,本例中输入 28,然后按 Enter 键。

因为前面选择的是层内插补,程序紧接着会询问分层变量的开始行列数、终止行列数,本例中作为分层变量的 C4,始于第 1 行的 17 列,终止于第 1 行的 20 列,见附图 2.39。然后屏幕会询问以及分层变量的层数,本例中分层变量 C4 的层数为 3,见附图 2.40。



附图 2.39



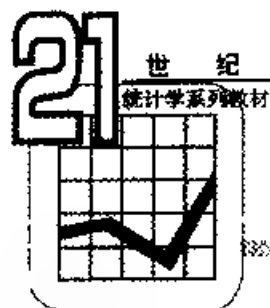
附图 2.40

处理数据过程中,PRE CARP 要求用户依次输入各层信号(1.2. 和 3.),插补后的结果见附图 2.41。原数据中的缺失值已被插补值代替,对每个观测单元都多出一个三维的指示向量,比如第一个观测单元的指示向量值为(0 1 1),说明该观测单元的 Y2 值是插补值,而 Y3 和 C6 的值为原始值。该数据集中每个观测单元的记录为 2 行,其数据格式为(2I2,6F4.0,/3F2.0) 其中的 / 表示后面 3 个 2 位的指示变量在第二行。

1 1 10.11.11. 2. 1. 1.
 0 1 1
 1 2 10.11.13. 2. 2. 2.
 1 1 1
 1 3 10 12.10. 1. 3. 1.
 1 1 0
 1 4 10.11. 7. 2. 1. 2.
 0 1 0
 1 5 10. 6. 5. 1. 2. 1.
 1 1 1
 1 5 10. 4. 5. 1. 3. 2.
 1 0 1
 2 1 20. 3. 6. 1. 2. 1.
 1 1 1
 2 2 20. 6.10. 2. 1. 1.
 1 1 1
 2 3 20.14.12. 2. 1. 1.
 1 1 1
 2 4 20. 6. 4. 1. 2. 1.
 1 1 1
 3 1 5. 1.15. 3. 4. 1.
 0 1 0
 3 1 5. 1.15. 3. 4. 1.
 1 0 0
 3 2 5. 2. 3. 3. 4. 1.
 1 1 1
 3 2 5. 3. 1. 3 4. 1.
 1 1 1
 3 2 5. 5. 6. 3. 4. 2.
 1 1 1
 4 1 4. 9. 7. 2. 1. 1.
 1 1 1
 4 1 4. 7. 4. 1. 2. 1.
 1 1 1

4	2	4	10	12	2	1	1	
					1	1	1	
4	2	4	15	14	2	2	2	
					1	1	1	
4	3	4	5	8	1	3	1	
					1	1	0	
4	3	4	7	8	1	2	1	
					1	1	0	

附图 2.41 插补后数据



附录 3

随机数表

03 47 43 73 86	36 96 47 36 61	46 98 63 71 62	33 26 16 80 45	60 11 14 10 95
97 74 24 67 62	42 81 14 57 20	42 53 32 37 32	27 07 36 07 51	24 51 79 89 73
16 76 62 27 66	56 50 26 71 07	32 90 79 78 53	13 55 38 58 59	88 97 54 14 10
12 56 85 99 26	96 96 68 27 31	05 03 72 93 15	57 12 10 14 21	88 26 49 81 76
55 59 56 35 64	38 54 82 46 22	31 62 43 09 90	06 18 44 32 53	23 83 01 30 30
16 22 77 94 39	49 54 43 54 82	17 37 93 23 78	87 35 20 96 43	84 26 34 91 64
84 42 17 53 31	57 24 55 06 88	77 04 74 47 67	21 76 33 50 25	83 92 12 06 76
62 01 63 78 59	16 95 55 67 19	98 10 50 71 75	12 86 73 58 07	44 39 52 38 79
33 21 12 34 29	78 64 56 07 82	52 42 07 44 38	15 51 00 13 42	99 66 02 79 54
57 60 86 32 44	09 47 27 96 54	49 17 46 09 62	90 52 84 77 27	08 02 73 43 28
18 18 07 92 45	44 17 16 58 09	79 83 86 19 62	06 76 50 03 10	55 23 64 05 05
26 62 38 97 75	84 16 07 44 99	83 11 46 32 24	20 14 85 88 45	10 93 72 88 71
23 42 40 64 74	82 97 77 77 81	07 45 32 14 08	32 98 94 07 72	93 85 79 10 75
52 36 28 19 95	50 92 26 11 97	00 56 76 31 38	80 22 02 53 53	86 60 42 04 53
37 85 94 35 12	83 39 50 08 30	42 34 07 96 88	54 42 06 87 98	35 85 29 48 39

70 29 17 12 13	40 33 20 38 26	13 89 51 03 74	17 76 37 13 04	07 74 21 19 30
56 62 18 37 35	96 83 50 87 75	97 12 25 93 47	70 33 24 03 54	97 77 46 44 80
99 49 57 22 77	88 42 95 45 72	16 64 36 16 00	04 43 18 66 79	94 77 24 21 90
16 08 15 04 72	33 27 14 34 09	45 59 34 68 49	12 72 07 34 45	99 27 72 95 14
31 16 93 32 43	50 27 89 87 19	20 15 37 00 49	52 85 66 60 44	38 68 88 11 80
68 34 30 13 70	55 74 30 77 40	44 22 78 84 26	04 33 46 09 52	68 07 97 06 57
74 57 25 65 76	59 29 97 68 60	71 91 38 67 54	13 58 18 24 76	15 54 55 95 52
27 42 37 86 53	48 55 90 65 72	96 57 69 36 10	96 46 92 42 45	97 60 49 04 91
00 39 68 29 61	66 37 32 20 30	77 84 57 03 29	10 45 65 04 26	11 04 96 67 24
29 94 98 94 24	68 49 69 10 82	53 75 91 93 30	34 25 20 57 27	40 48 73 51 92
16 90 82 66 59	83 62 64 11 12	67 19 00 71 74	60 47 21 29 68	02 02 37 03 31
11 27 94 75 06	06 09 19 74 66	02 94 37 34 02	76 70 90 30 86	38 45 94 30 38
35 24 10 16 20	33 32 51 26 38	79 78 45 04 91	16 92 53 56 16	02 75 50 95 98
38 23 16 86 38	42 38 97 01 50	87 75 66 81 41	40 01 74 91 62	48 51 84 08 32
31 96 25 91 47	96 44 33 49 13	34 86 82 53 91	00 52 43 48 85	27 55 26 89 62
00 07 40 67 14	64 05 71 95 80	11 05 65 09 68	76 83 20 37 90	57 16 00 11 66
14 90 84 45 11	75 73 88 05 90	52 27 41 14 86	22 98 12 22 08	07 52 74 95 80
68 05 51 18 00	33 96 02 75 19	07 60 62 93 55	59 33 82 43 90	49 37 38 44 59
20 46 78 73 90	97 51 40 14 02	04 02 33 31 08	39 54 16 49 36	47 95 93 13 30
64 19 58 97 79	15 06 15 93 20	01 90 10 75 06	40 78 78 89 62	02 67 74 17 33
05 26 93 70 60	22 35 85 15 13	92 03 51 59 77	59 56 78 06 83	52 91 05 70 74
07 97 10 88 23	09 98 42 99 64	61 71 62 99 15	06 51 29 16 93	58 05 77 09 51
68 71 86 85 82	54 87 66 47 54	73 32 08 11 12	44 95 92 63 16	29 56 24 29 48
26 99 61 65 53	58 37 78 80 70	42 10 50 67 42	32 17 55 85 74	94 44 67 16 94
14 65 52 68 75	87 59 36 22 41	26 78 63 06 55	13 08 27 01 50	15 29 39 39 43
17 53 77 58 71	71 41 61 50 72	12 41 94 96 26	44 95 27 36 99	02 96 74 30 83
90 26 59 21 19	23 52 23 33 12	96 93 02 18 39	07 02 18 36 07	25 99 32 70 23
41 23 52 55 99	31 04 49 69 96	10 47 48 45 88	13 41 43 89 20	97 17 14 49 17
60 20 50 81 69	31 99 73 68 68	35 81 33 03 76	24 30 12 48 60	18 99 10 72 34
91 25 38 05 90	94 58 28 41 36	45 37 59 03 09	90 35 57 29 12	82 62 54 65 60
34 50 57 74 37	98 80 33 00 91	09 77 93 19 82	74 94 80 04 04	45 07 31 66 49

85 22 04 39 43	73 81 53 94 79	33 62 46 86 28	08 31 54 46 31	53 94 13 38 47
09 79 13 77 48	73 82 97 22 21	05 03 27 24 83	72 89 44 05 60	35 80 39 94 88
88 75 80 18 14	22 95 75 42 49	39 32 82 22 49	02 48 07 70 37	16 04 61 67 87
90 96 23 70 00	39 00 03 06 90	55 85 78 38 36	94 37 30 69 32	90 89 00 76 33
53 74 23 99 67	61 32 28 69 84	94 62 67 86 24	98 33 41 19 95	47 53 53 38 09
63 38 06 86 54	99 00 65 26 94	02 82 90 23 07	79 62 67 80 60	75 91 12 81 19
35 30 58 21 46	06 72 17 10 94	25 21 31 75 86	49 28 24 00 49	55 65 79 78 07
63 43 36 82 69	65 51 18 37 88	61 38 44 12 45	32 92 85 88 65	54 34 81 85 35
98 25 37 55 26	01 91 82 81 46	74 71 12 94 97	24 02 71 37 07	03 92 18 66 75
02 63 21 17 69	71 50 80 89 56	38 15 70 11 48	43 40 45 86 98	00 83 26 91 03
64 55 22 21 82	48 22 28 06 00	61 54 13 43 91	82 78 12 23 29	06 66 24 12 27
85 07 26 13 89	01 10 07 82 04	59 63 69 36 03	69 11 15 83 80	13 29 54 19 28
58 54 16 24 15	51 54 44 82 00	62 61 65 04 69	38 18 65 18 97	85 72 13 49 21
34 85 27 84 87	61 48 64 56 26	90 18 48 13 26	37 70 15 42 57	65 65 80 39 07
03 92 18 27 46	57 99 16 96 56	30 33 72 85 22	84 64 38 56 98	99 01 30 98 64
62 93 30 27 59	37 75 41 66 48	86 97 80 61 45	23 53 04 01 63	45 76 08 64 27
08 45 93 15 22	60 21 75 46 91	98 77 27 85 42	28 88 61 08 84	69 62 03 42 73
07 08 55 18 40	45 44 75 13 90	24 94 96 61 02	57 55 66 83 15	73 42 37 11 61
01 85 89 95 66	51 10 19 34 88	15 84 97 19 75	12 76 39 43 78	64 63 91 08 25
72 84 71 14 35	19 11 58 49 26	50 11 17 17 76	86 31 57 20 18	95 60 78 46 75
88 78 28 16 84	13 52 53 94 53	75 45 69 30 96	73 89 65 70 31	99 17 43 48 76
45 17 75 65 57	28 40 19 72 12	25 12 74 75 67	60 40 60 81 19	24 62 01 61 16
96 76 28 12 54	22 01 11 94 25	71 96 16 16 88	68 64 36 74 45	19 59 50 88 92
43 31 67 72 30	24 02 94 08 63	38 32 36 66 02	69 36 38 25 39	48 03 45 15 22
50 44 66 44 21	66 06 58 05 62	68 15 54 35 02	42 35 48 96 32	14 52 41 52 48
22 66 22 15 86	26 63 75 41 99	58 42 36 72 24	58 37 52 18 51	03 37 18 39 11
96 24 40 14 51	23 22 30 88 57	95 67 47 29 83	94 69 40 06 07	18 16 36 78 86
31 73 91 61 19	60 20 72 93 48	98 57 07 23 69	65 95 39 69 58	56 80 30 19 44
78 60 73 99 84	43 89 94 36 45	56 69 47 07 41	90 22 91 07 12	18 35 34 08 72
84 37 90 61 56	70 10 23 98 05	85 11 34 76 60	76 48 45 34 60	01 64 18 39 96

36 67 10 08 23	98 93 35 08 86	99 29 76 29 81	33 34 91 58 93	63 14 52 32 52
07 28 59 07 48	89 64 58 89 75	83 85 62 27 89	30 14 78 56 27	86 63 59 80 02
10 15 83 87 60	79 24 31 66 56	21 48 24 06 93	91 98 94 05 49	01 47 59 38 00
55 19 68 97 65	03 73 52 16 56	00 53 55 90 27	33 42 29 38 87	22 13 88 83 34
53 81 29 13 39	35 01 20 71 34	62 33 74 82 14	53 73 19 09 03	56 54 29 56 93
51 86 32 68 92	33 98 74 66 99	40 14 71 94 58	45 94 19 38 81	14 44 99 81 07
35 91 70 29 13	80 03 54 07 27	96 94 78 32 66	50 95 52 74 33	13 80 55 62 54
37 71 67 95 13	20 02 44 95 94	64 85 04 05 72	01 32 90 76 14	53 89 74 60 41
93 66 13 83 27	92 79 64 64 72	28 54 96 53 84	48 14 52 98 94	56 07 93 89 30
02 96 08 45 65	13 05 00 41 84	93 07 54 72 59	21 45 57 09 77	19 48 56 27 44
49 83 43 48 35	82 88 33 69 96	72 36 04 19 76	47 45 15 18 60	82 11 08 95 97
84 60 71 62 46	40 80 81 30 37	34 39 23 05 38	25 15 35 71 30	88 12 57 21 77
18 17 30 88 71	44 91 14 88 47	89 23 30 63 15	56 34 20 47 89	99 82 93 24 98
79 69 10 61 78	71 32 76 95 62	87 00 22 58 40	92 54 01 75 25	43 11 71 99 31
75 93 36 57 83	56 20 14 82 11	74 21 97 90 65	96 42 68 63 86	74 54 13 26 94
38 30 92 29 03	06 28 81 39 38	62 25 06 84 63	61 29 08 93 67	04 32 92 08 09
51 29 50 10 34	31 57 75 95 80	51 97 02 74 77	76 15 48 49 44	18 55 63 77 09
21 31 38 86 24	37 79 81 53 74	73 24 16 10 33	52 83 90 94 76	70 47 14 54 36
29 01 23 87 88	58 02 39 37 67	42 10 14 20 92	16 55 23 42 45	54 96 09 11 06
95 33 95 22 00	18 74 72 00 18	38 79 58 69 32	81 76 80 26 92	82 80 84 25 39
90 84 60 79 80	24 36 59 87 38	82 07 53 89 35	96 35 23 79 18	05 98 90 07 35
46 40 62 98 82	54 97 20 56 95	15 74 80 08 32	16 46 70 50 80	67 72 16 42 79
20 31 89 03 43	38 46 82 60 72	32 14 82 99 70	80 60 47 18 97	63 49 30 21 30
71 59 73 05 50	08 22 23 71 77	91 01 93 20 49	82 96 59 26 94	66 39 67 98 60
22 17 68 65 84	68 95 23 92 35	87 02 22 57 51	61 09 43 95 06	58 24 82 03 47
19 36 27 59 46	13 79 93 37 55	39 77 32 77 09	85 52 05 30 30	47 83 51 62 74
16 77 23 02 77	09 61 87 25 21	28 06 24 25 93	16 71 13 59 78	23 05 47 47 25
78 43 76 71 61	20 44 90 32 64	97 67 63 99 61	46 38 03 93 22	69 81 21 99 21
03 28 28 26 08	73 37 32 04 05	60 30 16 09 05	88 69 58 28 99	35 07 44 75 47
93 22 53 64 39	07 10 63 76 35	87 03 04 79 88	08 13 13 85 51	55 34 57 72 69
78 76 58 54 74	92 38 70 96 92	52 06 79 79 45	82 63 18 27 44	69 66 92 19 09

23 68 35 26 00	99 53 93 61 28	52 73 05 48 34	56 65 05 61 86	90 92 10 70 80
15 39 25 70 99	93 86 52 77 65	15 33 59 05 28	22 87 26 07 47	86 96 98 29 06
58 71 96 30 24	18 46 23 34 27	85 13 99 24 44	49 18 09 79 49	74 16 32 23 02
57 35 27 33 72	24 53 63 94 09	41 10 76 47 91	44 04 95 49 66	39 60 04 59 81
48 50 86 54 48	22 06 34 72 52	82 21 15 65 20	33 29 94 71 11	15 91 29 12 03
61 96 48 95 03	07 16 39 33 66	98 56 10 56 79	77 21 30 27 12	90 49 22 23 62
36 93 89 41 23	29 70 83 63 51	99 74 20 52 36	87 09 41 15 09	98 60 16 03 03
18 87 00 42 31	57 90 12 02 07	23 47 37 17 31	54 08 01 88 63	39 41 88 92 10
88 56 53 27 59	33 35 72 67 47	77 34 55 45 70	08 18 27 38 90	16 95 86 70 75
09 72 95 84 29	49 41 31 06 70	42 38 06 45 18	64 84 73 31 65	52 53 37 97 15
12 96 88 17 31	65 19 69 02 83	60 75 86 90 68	24 64 19 35 51	56 61 87 39 12
85 94 57 24 16	92 09 84 38 76	22 00 27 69 85	29 81 94 78 70	21 94 47 90 12
38 64 43 59 98	98 77 87 68 07	91 51 67 62 44	40 98 05 93 78	23 32 65 41 18
53 44 09 42 72	00 41 86 79 79	68 47 22 00 20	35 55 31 51 51	00 83 63 22 55
40 76 66 26 84	57 99 99 90 37	36 63 32 08 58	37 40 13 68 97	87 64 81 07 83
02 17 79 18 05	12 59 52 57 02	22 07 90 47 03	28 14 11 30 79	20 69 22 40 98
95 17 82 06 53	31 51 10 96 46	92 06 88 07 77	56 11 50 81 69	40 23 72 51 39
35 76 22 42 92	96 11 83 44 80	34 68 35 48 77	33 42 40 90 60	73 96 53 97 86
26 29 13 56 41	85 47 04 66 08	34 72 57 59 13	82 43 80 46 15	38 26 61 70 04
77 80 20 75 82	72 82 32 99 90	63 95 73 76 63	89 73 44 99 05	48 67 26 43 18
46 40 66 44 52	91 36 74 43 53	30 82 13 54 00	78 45 63 98 35	55 03 36 67 68
37 56 08 18 09	77 53 84 46 47	31 91 18 95 58	24 16 74 11 53	44 10 13 85 57
61 65 61 68 66	37 27 47 39 19	84 83 70 07 48	53 21 40 06 71	95 06 79 88 54
93 43 69 64 07	34 18 04 52 35	56 27 09 24 86	61 85 53 83 45	19 90 79 99 00
21 96 60 12 99	11 20 99 45 18	48 13 93 55 34	18 37 79 49 90	65 97 38 20 46
95 20 47 97 97	27 37 83 28 71	00 06 41 41 74	45 89 09 39 84	51 67 11 52 49
97 86 21 78 73	10 65 81 92 59	58 76 17 14 97	04 76 62 16 17	17 95 70 45 80
69 92 06 34 13	59 71 74 17 32	27 55 10 24 19	28 71 82 13 74	63 52 52 01 41
04 31 17 21 56	33 73 99 19 87	26 72 39 27 67	53 77 57 68 93	60 61 97 22 61
61 06 98 03 91	87 14 77 43 96	43 00 65 98 50	45 60 33 01 07	98 99 46 50 47
85 93 85 86 88	72 87 08 62 40	16 06 10 89 20	23 21 34 74 97	76 38 03 29 63

21 74 32 47 45	73 96 07 94 52	09 65 90 77 47	25 76 16 19 33	53 05 79 53 30
15 69 53 82 80	79 96 23 53 10	65 39 07 16 29	45 33 02 43 79	02 87 40 41 45
02 89 08 04 49	20 21 14 68 86	87 63 93 95 17	11 29 01 95 80	35 14 97 35 33
87 18 15 89 79	85 43 01 72 73	08 61 74 51 69	89 74 39 82 15	94 51 33 41 67
98 83 71 94 22	59 97 50 99 52	08 52 85 08 40	87 80 61 65 31	91 51 80 32 44
10 08 58 21 66	72 68 49 29 31	89 85 84 46 06	59 73 19 85 23	65 09 29 75 63
47 90 56 10 08	88 02 84 27 83	42 29 72 23 19	66 56 45 65 79	20 71 53 20 25
22 85 61 68 90	49 64 92 85 44	16 40 12 89 88	50 14 49 81 06	01 82 77 45 12
67 80 43 79 33	12 83 11 41 16	25 58 19 68 70	77 02 54 00 52	53 43 37 15 26
27 62 50 96 72	79 44 61 40 15	14 53 40 65 39	27 31 58 50 28	11 39 03 34 25
33 78 80 87 15	38 30 06 38 21	14 47 47 07 26	54 96 87 53 32	40 36 40 96 76
13 13 92 66 99	47 24 49 57 74	32 25 43 62 17	10 97 11 69 84	99 63 22 32 98
10 27 53 96 23	71 50 54 36 23	54 31 04 82 98	04 14 12 15 09	26 78 25 47 47
28 41 50 61 88	64 85 27 20 18	83 36 36 05 56	39 71 65 09 62	94 76 62 11 89
34 21 42 57 02	59 19 18 97 48	80 30 03 30 98	05 24 67 70 07	84 97 50 87 40
61 81 77 23 23	82 82 11 54 08	53 28 70 58 96	44 07 39 55 43	42 34 43 39 28
61 15 18 13 54	16 86 20 26 88	90 74 80 55 09	14 53 90 51 17	52 01 63 01 59
91 76 21 64 64	44 91 13 32 97	75 31 62 66 54	84 80 32 75 77	56 08 25 70 29
00 97 79 08 06	37 30 28 59 85	53 56 68 53 40	01 74 39 59 73	30 19 99 85 48
36 46 18 34 94	75 20 80 27 77	78 91 69 16 00	08 43 18 73 69	67 69 61 34 25
88 98 99 60 50	65 95 79 42 94	93 62 40 89 96	43 56 47 71 66	46 76 29 67 02
04 37 59 87 21	05 02 03 24 17	47 97 81 56 51	92 34 86 01 82	55 51 33 12 91
63 62 06 34 41	94 21 78 55 09	72 76 45 16 94	29 95 81 83 83	79 88 01 97 30
78 47 23 53 90	34 41 92 45 71	09 23 70 70 07	12 38 92 79 43	14 85 11 47 23
87 68 62 15 43	53 14 36 59 25	54 47 33 70 15	59 24 48 40 35	50 03 42 99 36
47 60 92 10 77	88 59 53 11 52	66 25 69 07 04	48 68 64 71 06	61 65 70 22 12
56 88 87 59 41	65 28 04 67 53	95 79 88 37 31	50 41 06 94 76	81 83 17 16 33
02 57 45 86 67	73 43 07 34 48	44 26 87 93 29	77 09 61 67 84	06 69 44 77 75
31 54 14 13 17	48 62 11 90 60	68 12 93 64 28	46 24 79 16 76	14 60 25 51 01
28 50 16 43 36	28 97 85 58 99	67 22 52 76 23	24 70 36 54 54	59 28 61 71 96
63 29 62 66 50	02 63 45 52 38	67 63 47 54 75	83 24 78 43 20	92 63 13 47 48

45 65 58 26 51	76 96 59 38 72	86 57 45 71 46	44 67 76 14 55	44 88 01 62 12
39 65 36 63 70	77 45 85 50 51	74 13 39 35 22	30 53 36 02 95	49 34 88 73 61
73 71 98 16 04	29 18 94 51 23	76 51 94 84 86	79 93 96 38 63	08 58 25 58 94
72 20 56 20 11	72 65 71 08 86	79 57 95 13 91	97 48 72 66 48	09 71 17 24 89
75 17 26 99 76	89 37 20 70 01	77 31 61 95 46	26 97 05 73 51	53 33 18 72 87
7 48 60 82 29	81 30 15 39 14	48 38 75 93 29	06 87 37 78 48	45 56 00 84 47
68 08 02 80 72	83 71 46 30 49	89 17 95 88 29	02 39 56 03 46	97 74 06 56 17
14 23 98 61 67	70 52 85 01 50	01 84 02 78 43	10 62 98 19 41	18 83 99 47 99
49 08 96 21 44	25 27 99 41 28	07 41 08 34 66	19 42 74 39 91	41 96 53 78 72
78 37 06 08 43	63 61 62 42 29	39 68 95 10 96	09 24 23 00 62	56 12 80 73 16
37 21 34 17 68	68 96 83 23 56	32 84 60 15 31	44 73 57 34 77	91 15 79 74 58
14 29 09 34 04	87 83 07 55 07	76 58 30 83 64	87 29 25 58 84	86 50 60 00 25
58 43 28 06 36	49 52 83 51 14	47 56 91 29 34	05 87 31 06 95	12 45 57 09 09
10 43 67 29 70	80 62 80 03 42	10 80 21 38 84	90 56 35 03 09	43 12 74 49 14
44 38 88 39 54	86 97 37 44 22	00 95 01 31 76	17 16 29 56 63	38 78 94 49 81
90 69 59 19 51	85 39 52 85 13	07 28 37 07 61	11 16 36 27 03	78 86 72 04 95
41 47 10 25 62	97 05 31 03 61	20 26 36 31 62	68 69 86 95 44	84 95 48 46 45
91 94 14 63 19	75 89 11 47 11	31 56 34 19 09	79 57 92 36 59	14 93 87 81 40
80 06 54 18 66	09 18 94 06 19	98 40 07 17 81	22 45 44 84 11	24 62 20 42 31
67 72 77 63 48	84 08 31 55 58	24 33 45 77 58	80 45 67 93 82	75 70 16 08 24
59 40 24 13 27	79 26 88 86 30	01 31 60 10 39	53 58 47 70 93	85 81 56 39 38
05 90 35 89 95	01 61 16 96 94	50 78 13 69 36	37 68 53 37 31	71 26 35 03 71
44 43 80 69 98	46 68 05 14 82	90 78 50 05 62	77 79 13 57 44	59 60 10 39 66
61 81 31 96 82	00 57 25 60 59	46 72 60 18 77	55 66 12 62 11	08 99 55 64 57
42 88 07 10 05	24 98 65 63 21	47 21 61 88 32	27 80 30 21 60	10 92 35 36 12
77 94 30 05 39	28 10 99 00 27	12 73 73 99 12	49 99 57 94 82	96 88 57 17 91
78 83 19 76 16	94 11 68 84 26	23 54 20 86 85	23 86 66 99 07	36 37 34 92 09
87 76 59 61 81	43 63 63 61 61	65 76 36 95 90	18 48 27 45 68	27 23 65 30 72
91 43 05 96 47	55 78 99 95 24	37 55 85 78 78	01 48 41 19 10	35 19 54 07 73
84 97 77 72 73	09 62 06 65 72	87 12 49 03 60	41 15 20 76 27	50 47 02 29 16
87 41 60 76 83	44 88 96 07 80	83 05 83 38 96	73 70 66 81 90	30 56 10 48 59

28 89 65 87 08	13 50 63 04 23	25 47 57 91 13	52 62 24 19 94	91 67 48 57 10
30 29 43 65 42	78 66 28 55 80	47 46 41 90 08	55 98 78 10 70	49 92 05 12 07
95 74 62 60 53	51 57 32 22 27	12 72 72 27 77	44 67 32 23 13	67 95 07 76 30
01 85 54 96 72	66 86 65 64 60	56 59 75 36 75	46 44 33 63 71	54 50 06 44 75
10 91 46 96 86	19 83 52 47 53	65 00 51 93 51	30 80 05 19 29	56 23 27 19 03
05 33 18 08 51	51 78 57 26 17	34 87 96 23 95	89 9 93 39 79	11 28 04 15 52
04 43 13 37 00	79 68 96 26 60	70 39 83 66 56	62 03 55 86 57	77 55 33 62 02
05 85 40 25 24	73 52 93 70 50	48 21 47 74 63	17 27 27 51 26	35 96 29 00 45
84 90 90 65 77	63 99 25 69 02	09 04 03 35 79	19 79 95 07 21	02 84 48 51 97
28 55 53 09 48	86 28 30 02 35	71 30 32 06 47	93 74 21 86 33	49 90 21 69 74
89 83 40 69 80	97 96 47 59 97	56 33 24 87 36	17 18 16 90 46	75 27 28 52 13
73 20 96 05 68	93 41 69 96 07	97 50 81 79 59	42 37 13 81 83	92 42 85 04 31
10 89 07 76 21	40 24 74 36 42	40 33 04 46 24	35 63 02 31 61	34 59 43 36 96
91 50 27 78 37	06 06 16 25 98	17 78 80 36 85	26 41 77 63 37	71 63 94 94 33
03 45 44 66 88	97 81 26 03 89	39 46 67 21 17	98 10 39 33 15	61 63 00 25 92
89 41 58 91 63	65 99 59 97 84	90 14 79 61 55	56 16 88 87 60	32 15 99 67 43
13 43 00 97 26	16 91 21 32 41	60 22 66 72 17	31 85 33 69 07	68 49 20 43 29
71 71 00 51 72	62 03 89 26 32	35 27 99 18 25	78 12 03 09 70	50 93 19 35 56
19 28 15 00 41	92 27 73 40 38	37 11 05 75 16	98 81 99 37 29	92 20 32 39 67
56 37 39 82 39	45 51 94 69 04	00 84 14 36 37	95 66 39 01 09	21 68 40 95 79
39 27 52 89 11	00 81 06 28 48	12 08 05 75 26	03 35 63 05 77	13 81 20 67 58
73 13 28 58 01	05 06 42 24 07	60 60 29 99 93	72 93 78 04 36	25 76 01 54 03
81 60 84 51 57	12 68 46 55 89	60 09 71 87 89	70 81 10 95 91	83 79 68 20 66
05 62 98 07 85	07 79 26 69 61	67 85 72 37 41	85 79 76 48 23	61 58 87 08 05
62 97 16 29 18	52 16 16 23 56	62 95 80 97 63	32 25 34 03 36	48 84 60 37 65
31 13 63 21 08	16 01 92 58 21	48 79 74 73 72	08 64 80 91 38	07 28 66 61 59
97 38 35 34 19	89 84 05 34 47	88 09 31 54 88	97 96 86 01 69	46 13 95 65 96
32 11 78 33 82	51 99 98 44 39	12 75 10 60 36	80 66 39 94 97	42 36 31 16 59
81 99 13 37 05	08 12 60 39 23	61 73 84 89 18	26 02 04 37 95	96 18 69 06 30
45 74 00 03 05	69 99 47 26 52	48 06 30 00 18	03 30 28 55 59	66 10 71 44 05
11 84 13 69 01	88 91 28 79 50	71 42 14 96 55	98 59 96 01 36	88 77 90 45 59

14 66 12 87 22	59 45 27 08 51	85 64 23 85 41	64 72 08 59 44	67 98 36 65 56
40 25 67 87 82	87 27 17 30 37	48 69 49 02 58	98 02 50 58 11	95 39 06 35 63
44 48 97 49 43	65 45 53 41 07	14 83 46 74 11	76 66 63 60 08	90 54 33 65 84
41 94 54 06 57	48 28 01 83 84	09 11 21 91 73	97 28 44 74 06	22 30 95 69 72
07 12 15 58 84	93 18 31 83 45	54 52 62 29 91	53 58 54 66 05	47 19 63 92 75
64 27 90 43 52	18 26 32 96 83	50 58 45 27 57	14 96 39 64 85	73 87 96 76 23
80 71 86 41 03	45 62 63 40 88	35 69 34 10 94	32 22 52 04 74	69 63 21 83 41
27 06 08 09 92	26 22 59 28 27	38 58 22 14 79	24 32 12 38 42	33 56 90 92 57
54 68 97 20 54	33 26 74 03 30	74 22 19 13 48	30 28 01 92 49	58 61 52 27 03
02 92 65 68 99	05 53 15 26 70	04 69 22 64 07	04 73 25 74 82	78 35 22 21 88
83 52 57 78 62	98 61 70 48 22	68 50 64 55 75	42 70 32 09 60	58 70 61 43 97
82 82 76 31 33	85 13 41 38 10	16 47 61 43 77	83 27 19 70 41	34 78 77 60 25
38 61 34 09 49	04 41 66 09 76	20 50 73 40 95	24 77 95 73 20	47 42 80 61 03
01 01 11 88 38	03 10 16 82 24	39 58 20 12 39	82 77 02 18 88	33 11 49 15 16
21 66 14 38 28	54 08 18 07 04	92 17 63 36 75	33 14 11 11 78	97 30 53 62 33
32 29 30 69 59	68 50 33 31 47	15 64 88 75 27	04 51 41 61 96	86 62 93 66 71
04 59 21 65 47	39 90 89 86 77	46 86 86 88 86	50 09 13 24 91	54 80 67 78 66
38 64 50 07 36	56 50 45 94 25	48 28 48 30 51	60 73 73 03 87	68 47 37 10 84
48 33 50 83 53	59 77 64 59 90	58 92 62 50 18	93 09 45 89 06	13 26 98 86 29



习题参考答案

第2章

4. $\bar{Y} = 3, \sigma^2 = 5.2, S^2 = 6.5$
5. $\hat{Y} = 613\,800(\text{元}), s(\hat{Y}) = 79\,034(\text{元}), n = 465$
6. $p = 93\%, s(p) = 2.54\%, n = 1\,668$
7. $\hat{Y} = 20.7(\text{分钟}),$ 估计 95% 的置信区间 $(18.40, 23.00)$
8. $p = 6\%, s(p) = 3.4\%, n = 2\,121$

第3章

3. (1) $\bar{y}_q = 20.07(\text{元}), s(y_q) = 3.08(\text{元})$
 (2) 按比例分配 $n = 186, n_1 = 57, n_2 = 92, n_3 = 37$
 Neyman 分配 $n = 175, n_1 = 33, n_2 = 99, n_3 = 43$
4. (1) $p_q = 92.4\%, s(p_q) = 1.99\%$
 (2) 按比例分配 $n = 2\,663,$ 各层样本量为: 479, 559, 373, 240, 426, 586
 Neyman 分配 $n = 2\,565,$ 各层样本量为: 536, 520, 417, 304, 396, 392
5. $\bar{y}_t = 75.79(\text{元}),$ 置信区间 $(60.63, 90.95)$
6. $n = 92$

7. (1) 错 (2) 错 (3) 错 (4) 对 (5) 样本量足够大时是对的

8. (1) $p = 3\%$, $s(p) = 1.71\%$; (2) $p_{psr} = 2.68\%$, $s(p_{psr}) = 1.64\%$

第4章

4. $\hat{R} = 1.0468(\text{元})$, $s(\hat{R}) = 0.0900(\text{元})$

5. (1) $\hat{Y}_R = 231\,611.86(\text{元})$, $s(\hat{Y}_R) = 1\,536.92(\text{元})$;

$\hat{Y}_r = 231\,581.66(\text{元})$, $s(\hat{Y}_r) = 1\,475.42(\text{元})$

(2) 比率估计: (5.707%, 8.494%); 回归估计: (5.750%, 8.424%)

(3) 对于比估计, $n = 13$; 对于回归估计, $n = 12$

6. $x^{(1)}$, $x^{(2)}$ 均可, $x^{(1)}$ 更好

7. (2) 偏倚分别为: $-0.0274, 0.0040$, MSE 分别为: $0.1221, 0.0686$

第5章

5. $\hat{Y} = 2\,217.01$, $s(\hat{Y}) = 142.544$

7. $MSE(\hat{y}) = 11.5$, $MSE(\hat{Y}_R) = 0.9613$, $MSE(\hat{Y}_{HH}) = 0.5897$

第6章

2. (1) $\bar{y} = 19.73$, $v(\bar{y}) = 0.79$

(2) $\hat{Y} = N\bar{M}\bar{y} = 12\,311.52$, $v(\hat{Y}) = 554.62^2$

(3) $\hat{Y} = M_0\bar{y} = 14\,008.3$, $v(\hat{y}) = M_0^2 v(\bar{y})$ (4) $n = 14$

3. $\bar{y} = 1.875$, $\hat{Y} = 7\,500$, $v(\bar{y}) = 0.0089$, $v(\hat{Y}) = 141\,900$

4. $p = 0.4$, $v(p) = 0.0034$

5. (1) $p = 0.7091$, $s(p) = 0.0241$ (2) $n = 7$

6. $\hat{Y} = 3\,532.8$, $2\sqrt{v(\hat{Y})} = 539.50$

7. $\bar{y} = 5.91$, $2\sqrt{v(\bar{y})} = 0.322$

8. $\hat{Y} = 495\,392.4(\text{吨})$, $2\sqrt{v(\hat{Y})} = 19\,473$

9. (1) $t = 0.5$, $M = 5$ (2) $t = 2$, $M = 1$

第7章

3. (1) 样本单元: 5, 11, 17, 23, 29, 35, 1

(2) Sethi 对称系统样本单元: 5, 6, 15, 16, 25, 26, 35

Singn 对称系统样本单元: 5, 31, 10, 26, 15, 21, 20

4. 简单随机抽样: $V(y_{sr}) = \frac{N-n}{Nn} S^2 = 0.0034$

等距抽样: $V(y_{sy}) = \frac{1}{k} \sum_{r=1}^k (y_r - Y)^2 = 0.00141$

$V(y_{sy}) < V(y_{sr})$

5. 估计汉族所占的比例, 采用等距抽样效果最好

6. (1) 估计男性所占比例:

简单随机抽样方差 $V(\tilde{y}_{sr}) = \frac{N-n}{Nn} S^2 = 0.0204$

等距抽样方差 $V(y_{sy}) = \frac{1}{k} \sum_{r=1}^k (y_r - \bar{Y})^2 = 0.0216$

$V(y_{sy}) > V(y_{sr})$

(2) 估计孩子所占比例:

简单随机抽样方差 $V(y_{sr}) = \frac{N-n}{Nn} S^2 = 0.0204$

等距抽样方差 $V(y_{sy}) = \frac{1}{k} \sum_{r=1}^k (y_r - \bar{Y})^2 = 0.0776$

$V(y_{sy}) > V(y_{sr})$

(3) 估计具有某种职业的住户中人员的比例:

简单随机抽样方差 $V(\tilde{y}_{sr}) = \frac{N-n}{Nn} S^2 = 0.01923$

等距抽样方差 $V(\tilde{y}_{sy}) = \frac{1}{k} \sum_{r=1}^k (y_r - Y)^2 = 0.0016$

$V(y_{sy}) < V(y_{sr})$

7. 简单随机抽样方差 $V(y_{sr}) = 5.33$,

等距抽样方差 $V(y_{sy}) = 2$, $V(y_{sy}) < V(\tilde{y}_{sr})$

8. (1) 书稿的平均错字数: $y_{\bar{y}} = 4.7333$

(2) 用合并层方法估计抽样方差:

$v_3 = \frac{1-f}{n} \cdot \frac{2}{n} \cdot \frac{1}{2} \sum_{i=1}^{n/2} (y_{2i} - y_{2i-1})^2 = 0.131556$

(3) 用连续差方法估计抽样方差:

$v_2 = \frac{1-f}{n} \cdot \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 = 0.167356$

(4) 用交叉子样本法估计抽样方差: 多解, 如 $m = 2$,

$v_4 = \frac{1}{m(m-1)} \sum_{\alpha=1}^m (y_{\alpha} - \hat{Y})^2 = 0.04$

第8章

3. $p = 0.3, s(p) = 0.077$

4. $m = 2, n = 20$

5. (1) $\bar{y} = 3.3, s(\bar{y}) = 0.4043$; (2) $\bar{y} = 3.9521, s(\bar{y}) = 0.2674$

6. $\bar{y} = 34, s(\bar{y}) = 6$

第9章

1. 全县棉花种植面积估计为: $\hat{Y} = N\bar{y}_{stD} = 2000 \times 164.27 = 328540$

\hat{Y} 抽样标准误估计为: $s(\hat{Y}) = Ns(y_{stD}) = 38289.68$

2. (1) 二重抽样中最优的 $n_1 = n_2 = 268, n' = 1271$

二重分层抽样方差估计为: $v(p_{RD}) = 6.67$

(2) 不分层的简单随机抽样方差 $v_{sr} = 8.33$, 故二重抽样效率高

(3) $\frac{c_2}{c_1} > 9$ 时, 二重抽样的效率高于简单随机抽样

3. (1) 用二重比估计法估计该地区年末牛的总头数为: $\hat{Y} = 745713.2$

\hat{Y} 抽样标准误为: $s(\hat{Y}) = Ns(y_{stD}) = 1238 \times \sqrt{1404.583}$
 $= 46397.48$

(2) 使用二重回归估计法估计

$y_{brD} = y + b(x' - \bar{x}) = 599.5996$

该地区年末牛的总头数 $\hat{Y} = 742304.3$

\hat{Y} 抽样标准误为: $s(\hat{Y}) = Ns(y_{stD}) = 1238 \times \sqrt{1331.031}$
 $= 45166.32$

4. 相对于 n 来说, n' 必须大于 $26n$

5. Y 的二重回归估计量的标准差为 1.05

6. (1) 如果 $c_1 = \frac{c_{2h}}{100}$, 二重抽样的样本最优分配方案:

$f_1 = 0.133, f_2 = 0.229, n' = 612, n_1 = 64, n_2 = 30$

此时, $v(y_{stD}) = 4.71$

(2) $\frac{c_1}{c_{2h}} < 0.11$, 二重抽样的精度高于简单随机抽样

7. $S^2 \approx \sum W_h S_h^2 + \sum W_h (Y_h - Y)^2 = \frac{1}{L} \sum_h S_h^2 + \sum \frac{(\bar{Y}_h - Y)^2}{L}$

代入 y_{stD} 的方差公式即可

$$V(y_{SD}) = \left(\frac{1}{n} - \frac{1}{N} \right) S^2 + \sum_{h=1}^L \frac{W_h S_h^2}{n} \left(\frac{1}{f_{hL}} - 1 \right)$$

第 10 章

1. 该镇失业率估计 $\hat{R} = 0.0903$

用随机组法进行方差估计 $v_1(\hat{R}) = 0.000007$

2. 学生对学校伙食的满意比率 \hat{R} 的估计为: $\hat{R} = 0.503464$

$$\hat{R} \text{ 的方差估计 } v_k(\hat{R}) = \frac{1}{4} \sum_{a=1}^4 (\hat{R}_a - \hat{R})^2 = 0.000189$$

3. 该镇人口出生率估计 $\hat{R} = 0.00958$

用刀切法进行方差估计 $v_1(\hat{R}) = 5.85 \times 10^{-7}$

第 11 章

4. 该总体真实均值为 $Y = 95\% \times 45.4 + 5\% \times 59.0 = 46.08$

(1) 对于一个在 60% 层中抽样的方法:

$$bias = 40.7 - 46.08 = -5.38$$

$$V(y) = \frac{100^2(1-p)p}{n} = \frac{40.7(100-40.7)}{n} = \frac{2414}{n}$$

$$MSE(y) = V(y) + bias^2 = \frac{2414}{n} + 28.94$$

(2) 当回答率为 60% 时, 由 (1) 有 $\sqrt{MSE(y)} > \sqrt{28.94} > 5$
即均方误差的根不可能达到 5%

当回答率为 80% 时, $bias = 43.5 - 46.08 = -2.58$

当回答率高于 80% 时, $bias < 2.58$

而对于所有的回答率方法均有 $V(y) \leq \frac{2500}{n}$

因而当采用 80% 或更高回答率时, $\sqrt{MSE(y)} < \sqrt{\frac{2500}{n} + 2.58^2}$

只要当 n 稍稍大于 100, 便有 $\sqrt{MSE(y)} < 5$

(3) 用 90% 方法时, $bias = 44.8 - 46.08 = -1.28$

$$\sqrt{MSE(y)} = \sqrt{V(y) + bias^2} = \sqrt{\frac{44.8 \times 55.2}{n} + (-1.28)^2} = 2$$

得 $n = 1047$

采用用 95% 方法时, $bias = 45.4 - 46.08 = -0.68$

$$\sqrt{MSE(\bar{y})} = \sqrt{V(\bar{y}) + bias^2} = \sqrt{\frac{45.4 \times 54.6}{n} + (-0.68)^2} = 2$$

得 $n = 701$

5. 由上题(3) 知, 当回答率为 90% 时, $n = 1\,047$

总费用 $= 5(1\,047) = 5\,235$

当回答率为 95% 时, $n = 701$

总费用 $= 701\left(\frac{90\%}{95\%}\right)5 + 701\left(\frac{5\%}{95\%}\right)20 = 4\,058$



参考文献

1. W. G. Cochran. 抽样技术. 北京: 中国统计出版社, 1985
2. 冯士雍, 施锡铨. 抽样调查——理论、方法与实践. 上海: 上海科学技术出版社, 1996
3. 冯士雍, 倪加勋, 邹国华. 抽样调查理论与方法. 北京: 中国统计出版社, 1998
4. 倪加勋. 抽样调查. 大连: 东北财经大学出版社, 1994
5. 倪加勋. 抽样技术习题解答. 北京: 中国统计出版社, 1992
6. 金勇进. 非抽样误差分析. 北京: 中国统计出版社, 1996
7. 加拿大统计局. 调查技能教程. 北京: 国家统计局统计教育中心、国际合作司, 2001
8. Kirk M. Wolter. 方差估计引论. 北京: 中国统计出版社, 1998
9. Judith T. Lessler, William D. Kalsbeek. 调查中的非抽样误差. 北京: 中国统计出版社, 1998
10. 李金昌. 抽样调查与推断. 北京: 中国统计出版社, 1996
11. 谢邦昌. 抽样调查的理论及其应用方法. 北京: 中国统计出版社, 1998
12. 樊鸿康. 抽样调查. 北京: 高等教育出版社, 2000
13. 黄良文, 吴国培. 应用抽样方法. 北京: 中国统计出版社, 1991

14. 施锡铨. 抽样调查的理论和方法. 上海: 上海财经大学出版社, 1996
15. 肖红叶, 周恒彤. 抽样调查设计原理. 北京: 经济科学出版社, 1997
16. 卢宗辉. 抽样方法的系统研究. 北京: 中国统计出版社, 1998
17. 王国明, 李学增, 刘晓越, 王文颖编译. 抽样原理及其应用. 北京: 中国统计出版社, 1996
18. L. Kish. 抽样调查. 北京: 中国统计出版社, 1997
19. 柯惠新, 刘红鹰. 民意调查实务. 北京: 中国经济出版社, 1996
20. 胡健颖, 孙山泽. 抽样调查的理论方法和应用. 北京: 北京大学出版社, 2000
21. 梁小筠, 祝大平. 抽样调查的方法和原理. 上海: 华东师范大学出版社, 1994
22. Bureau of Labor Statistics. U. S. Census Bureau 2000 Design and Methodology. www.bls.census.gov/cps
23. Wayne A. Fuller. PC CARP. Statistical Laboratory, Iowa State University, 1989
24. Anich. D. and Bettin. P. Choosing a Composite Estimator for CPS. presented for presentation at the International Symposium on Panel Surveys, Washington, DC, 1986
25. Fay. R. and Train. G. Aspects of Survey and Model – Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties. Proceedings of the Section on Government Statistics, American Statistical Association PP. 154 – 159, 1995
26. Reeder. Regional Response to Questions on CPS Type A Rates. Bureau of the Census, CPS office Memorandum No. 97 – 07, Methods and Performance Evaluation Memorandum No. 97 – 03, January 31, 1997
27. Brick. J. M. and G. Kalton. Handling Missing Data in Survey Research. Statistical Mathematics in Medical Research. 5: 215 – 238. , 1996
28. Cox. B. G. Business Survey Methods. John Wiley and Sons, New York, 1995
29. Dolson. D. Imputation Methods. Statistics Canada, 1999
30. Rousseeuw. P. J. and A. M. Leroy. Robust Regression and Outlier Detection. John Wiley and Sons, New York, 1987
31. R. L. Scheaffer, W. Mendenhall, L. Ott. Elementary Survey Sampling. PWS – KENT Publishing Company, 1990
32. United Nations. Recommendations Concerning the Preparation of Report on Sampling Survey. Statistical Papers Series C. No. 1, New York, revised in 1964,

Statistical Papers Series C, No. 1, rev. 2

33. Barnett, V. ,and T. Lewis. Outliers in Statistical Data. John Wiley and Sons, Chichester, 1995
34. Lee, H. Outliers in Business Survey. Business Survey Methods, John Wiley and Sons. New York. 503 - 526